EPIGENOMIC SIGNATURES OF NEURONAL DIVERSITY IN THE MAMMALIAN BRAIN

By

Alisa Mo

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

April 2015

© 2015 Alisa Mo
All rights reserved

Abstract

Neuronal diversity is essential for mammalian brain function but poses a challenge to molecular profiling. To address the need for tools that facilitate cell type-specific epigenomic studies, we developed the first affinity purification approach to isolate nuclei from genetically-defined cell types in a mammal. We combine this technique with next generation sequencing to show that three subtypes of neocortical neurons have highly distinctive epigenomic landscapes. Over 200,000 regions differ in chromatin accessibility and DNA methylation signatures characteristic of gene regulatory regions. By footprinting and motif analyses, these regions are predicted to bind distinct cohorts of neuron subtype-specific transcription factors. Neuronal epigenomes reflect both past and present gene expression, with DNA hyper-methylation at developmentally critical genes appearing as a novel epigenomic signature in mature neurons. Taken together, our findings link the functional and transcriptional complexity of neurons to their underlying epigenomic diversity.

Thesis advisor:

Jeremy Nathans, M.D., Ph.D.

Professor

Department of Molecular Biology and Genetics

Johns Hopkins University School of Medicine

Thesis reader:

Seth Blackshaw, Ph.D.

Associate Professor

The Solomon H. Snyder Department of Neuroscience

Johns Hopkins University School of Medicine

Acknowledgements

I owe my deepest gratitude to my advisor, Dr. Jeremy Nathans, for the privilege of working with him and in his lab. I have greatly benefited from his guidance, patience, and trust in me. Jeremy gave me tremendous freedom to explore my own ideas both at the bench and at the computer, yet he was always available when I needed help. From Jeremy, I learned that doing good science is a combination of hard work, creativity, forensic scrutiny of the data, and gaining a broad scientific knowledge base. Jeremy inspires me not only because he is a great scientist, but also because he is a person of uncommon generosity and integrity.

In the Nathans lab, I found a remarkable group of talented colleagues: Amir Rattner, Yanshu Wang, John Williams, Phil Smallwood, Hugh Cahill, Hao Wu, Hao Chang, Max Tischfield, Lucas Hua, Huimin Yu, Danfeng Cai, Yulian Zhou, Francesco Emiliani, Chris Cho, Mark Sabbagh, and Andy Gu. In addition to the daily scientific conversations and technical advice, what counts to me are the friendships we forged through our long hours in lab and our shared joys, laughs, and frustrations. I am particularly indebted to Max Tischfield and Hao Wu for their mentorship, scientific criticisms, and career advice during my first years in lab and our continued friendship.

I feel incredibly fortunate to have collaborated with the teams of Dr. Joe Ecker (Salk) and Dr. Sean Eddy (Janelia), in particular with Eran Mukamel (Salk/UCSD), Chongyuan Luo (Salk), Ryan Lister (Salk/UWA), Lee Henry (Janelia), and Fred Davis (Janelia). I am especially grateful to Joe for his enthusiasm and support over the years and for many vibrant scientific discussions. I have learned much about research from working with Eran, who contributed to the data analysis in this project and whose mentorship has been an important academic influence on me. I have also benefited greatly from working with Lee, who helped me develop the affinity purification strategy, and Fred and Chongyuan, who played critical roles in the data analysis. All of these folks not only contributed to the scientific work included in this thesis but also took the time to teach me an enormous amount of genomics, statistics, and computational analysis. I admired their patience when I peppered them with questions, and their honest critiques of my own data analysis and manuscript drafts always motivated me to up my game.

I would like to acknowledge Dr. Sol Snyder and Dr. David Ginty for giving me the opportunities to rotate in their labs. Through these rotations, I gained valuable perspectives on diverse topics within neuroscience. I am also grateful to each of my past and present thesis committee members for their insightful critiques which have helped mold my project: Dr. Seth Blackshaw, Dr. Hongjun Song, Dr. Sarah Wheelan, Dr. David Ginty, and Dr. Jef Boeke.

Johns Hopkins has a beautifully collegial environment, and I have been fortunate to work with professors, post-docs, and students from many labs who taught me new experimental techniques and who generously shared their reagents. I owe a deep thank you to Dr. Bob Siliciano and Dr. Andrea Cox in the MSTP program and to my pre-thesis advisor, Dr. Mike Caterina, for their continued encouragement over the course of my graduate training. In my attempts to navigate graduate school, I am grateful to Sharon Welling and Bern Harper in the MSTP office as well as Rita Ragan and Beth Wood-Roig in the neuroscience graduate program for always being friendly and helpful.

I am also indebted to my first mentors in science and medicine: Dr. Q. Richard Lu (UTSW), Dr. Owen Obel (Dallas VA), and Dr. Robert F. Gilmour Jr. (Cornell). In particular, working with Dr. Gilmour, together with colleagues Dr. Niels Otani and Dr. Flavio Fenton, helped give me a quantitative foundation to delve into the heart of science (pun intended).

My friends from my medical school class (med12!), the MSTP program, and the neuroscience graduate program have helped keep me sane and happy in Baltimore. Any list of names here would be woefully incomplete, but I owe special thanks to Phyllis Yang, Eunice

Yang, Hamid Alai, and Lisa Eisler for their wonderful friendship and support, especially during times when I felt discouraged with myself.

And finally, to my parents Li Mo and Yanan Long, thank you. This has been a long road for you too, and there's so much to thank you for. Without your love and guidance, I would never have come this far.

Table of Contents

Abstract.......... ii

Acknowledgements.......... iv

Table of Contents.......... vii

List of Figures.......... ix

Chapter I: Introduction.......... 1

Chapter II: Materials and Methods.......... 4

- A. Experimental Procedures.......... 4
- B. General Data Analysis.......... 10
- C. Figure-Specific Data Analysis.......... 17

Chapter III: Results.......... 26

- A. Mammalian INTACT Isolates Specific Populations of Neuronal Nuclei from the Brain.......... 26
- B. INTACT RNA-Seq Captures Neuronal Subtype Markers.......... 27
- C. Non-CG Methylation is a Common Feature of Both Excitatory and Inhibitory Neurons, but Shows Widespread Differences in Genomic Distribution.......... 28
- D. Neuronal Regulatory DNA is Predominantly Cell Type-Specific.......... 30
- E. Cell Type-Specific Hypo-Methylation at Activity-Induced Transcription Factor Binding Sites.......... 31
- F. Neuronal Subtypes Show Coordinated Epigenomic Differences.......... 32
- G. Distinct Sets of DNA Binding Factors Act at Putative Neuron Subtype-Specific Regulatory Regions.......... 34
- H. Among DNA Methylation and Chromatin Accessibility Features, Non-CG

Methylation Best Correlates With RNA Abundance.. 36
I. Gene Clusters Based on Intragenic Non-CG Methylation Share Gene Expression, Chromatin, and Functional Organization.. 36
J. Non-CG Methylation is Lowest at the Nucleosome Center and Increases at Linker Regions... 38
K. Identification of Distinct Classes of Large Hypo-Methylated Regions..................... 39
L. Hyper-Methylation at Cell Type-Specific Transcription Factor Genes Preserves a Trace of Early Developmental Expression.. 40
Chapter IV: Discussion.. 43
References.. 48
Figures.. 58
Curriculum Vitae.. 88

List of Figures

Figure 1. An Affinity Purification Method Isolates Cell Type-Specific Nuclei in Mice............... 58

Figure 2. Nuclear Labeling and Specificity of Mouse INTACT.. 60

Figure 3. Widespread Differences in Gene Expression and DNA Methylation Among Neuron Subtypes.. 63

Figure 4. Gene Expression and DNA Methylation Analysis.. 65

Figure 5. Epigenomic Marks are Coordinated and Highly Cell Type-Specific............................ 68

Figure 6. Correlations Across Epigenomic Marks and Relevance of Neuron Subtype-Specific Hypo-Methylation to Induced Neuronal Activity.. 70

Figure 7. Relationships Across Cell Types and Development Via Epigenomic Marks................ 73

Figure 8. Epigenomic Correlations Across Cell Types and Development................................... 74

Figure 9. Neuronal Subtypes are Associated with Distinct Patterns of TF Binding...................... 75

Figure 10. Analysis of Putative TF Binding at Neuronal Regulatory Regions............................ 77

Figure 11. Integrative Analysis of DNA Methylation, Gene Expression, and Chromatin Features... 79

Figure 12. Integrative Analysis of Epigenomic Features.. 81

Figure 13. Large Domains of Low Methylation Link to Gene Expression, Including Unexpected Hyper-Methylation at Developmental Genes... 83

Figure 14. Large Hypo-Methylated Domains.. 85

Chapter I

Introduction

In the mammalian brain, distinct types of neurons interact in intricate networks to govern thought, emotion, and behavior. Neurons can differ in their morphologies, synaptic connections, electrophysiological properties, neurotransmitter identities, and developmental histories. The balance of signaling across heterogeneous neurons is critical for healthy brain function, and disruptions of genes that mediate this balance are implicated in neurological and psychiatric diseases (Sullivan et al., 2012).

Neuronal diversity arises partly through spatiotemporal regulation of gene expression by regulatory regions such as promoters and enhancers. These discrete regions of DNA can be identified using epigenomic signatures, which include accessible chromatin, active histone modifications, and low levels of DNA methylation (Bird, 2002; Heintzman et al., 2007; Stadler et al., 2011; Thurman et al., 2012). Neurons undergo extensive epigenomic changes during post-natal brain development, including *de novo* establishment of non-CG methylation (Lister et al., 2013; Xie et al., 2012). However, the genome-wide patterns of accessible chromatin and both CG and non-CG methylation in specific neuronal subpopulations are unknown. We reasoned that neuronal epigenomic landscapes should mirror neuronal diversity. Whereas gene expression analysis provides a snapshot of a neuron's molecular activity at a single point in time, the complementary epigenomic information captures gene regulatory mechanisms, developmental origins, and potential future responses induced by neuronal activity.

Cellular diversity is important for brain function, but it also poses a technical challenge for epigenomic studies. Cell type-specific molecular profiling requires the isolation of targeted cell populations from complex tissues (Maze et al., 2014). Manual sorting (Sugino et al., 2006) and laser capture microdissection (Emmert-Buck et al., 1996) are useful for isolating small numbers of cells but do not provide enough material for epigenomic studies. Fluorescence-activated cell sorting (FACS) can isolate larger numbers of cells but may be challenging in tissues such as the adult brain, where cells are morphologically complex and densely interconnected. Although improvements have been made (Saxena et al., 2012), the neuronal dissociation process may also induce cellular stress responses and perturb subsequent molecular profiles. Genetically-directed strategies that isolate RNA from specific cell populations in mice (Doyle et al., 2008; Gay et al., 2013; Heiman et al., 2008; Sanz et al., 2009) have begun to chart transcriptional diversity across cell types but cannot profile epigenomic features unless combined with FACS (Mellén et al., 2012). Although nuclei can be isolated by FACS for epigenomic studies (Jiang et al., 2008), FACS-sorted nuclei are fragile and difficult to concentrate into the small volumes that are optimal for chromatin assays. An alternate approach is INTACT (isolation of nuclei tagged in specific cell types; Deal and Henikoff, 2010), which uses affinity purification to isolate tagged nuclei. Captured nuclei can be used for gene expression, epigenomic, and proteomic profiling (Amin et al., 2014; Henry et al., 2012; Steiner et al., 2012).

Here, we present the first application of INTACT in a mammalian organism and use it to address the cell type-specific neuronal epigenome. Our approach uses the Cre-loxP system in mice to express a tagged nuclear membrane protein, allowing affinity purification of labeled nuclei from genetically-defined cell populations. In this study, we applied INTACT to examine the core transcriptional and epigenomic features of three major functional classes of neocortical neurons: excitatory pyramidal neurons, *Parvalbumin*-expressing fast-spiking interneurons (PV), and *Vasoactive intestinal peptide*-expressing interneurons (VIP). 70-85% of cortical neurons are excitatory. The remaining 15-30% are inhibitory neurons, with approximately 40% expressing PV and 12% expressing VIP (Gelman and Marín, 2010; Rudy et al., 2011). Together, these mutually exclusive cell types represent both glutamatergic (excitatory) and GABAergic (inhibitory) signaling. Neocortical pyramidal neurons provide the long-range excitatory output of the brain, and inhibitory neurons modulate the rate and temporal structure of this network output (Molyneaux et al., 2007; Rudy et al., 2011). PV and VIP neurons have distinct computational roles as a result of differences in their firing patterns and synaptic connections (Kepecs and Fishell, 2014).

Several studies have identified genome-wide differences in gene expression across neuronal subpopulations (Doyle et al., 2008; Molyneaux et al., 2014; Sugino et al., 2006). However, neuron subtype-specific epigenomes remain largely unexplored. We find that among excitatory, PV, and VIP neurons, global epigenomic landscapes of DNA methylation and chromatin accessibility show widespread differences. These differences reflect distinct mechanisms of gene regulation, with candidate regulators identified using transcription factor (TF) footprinting and motif analyses. Integrating epigenomes together with expression profiles, we find intragenic non-CG methylation to be particularly salient for inferring neuronal gene expression. At TF genes with cell type-specific developmental roles, we further identify a unique pattern of DNA hyper-methylation in adult neurons that is a long-lasting epigenomic signature of transient expression during brain development. Collectively, our results provide a comprehensive view of how distinct neuronal classes adopt unique epigenomic and gene regulatory configurations that reflect both mature neuronal function as well as developmental origin.

Chapter II

Materials and Methods

A. Experimental Procedures

Mouse INTACT

Animal procedures were conducted in accordance with the Institutional Animal Care and Use Committee guidelines of the Johns Hopkins Medical Institutions. The *R26-CAG-LSL-Sun1-sfGFP-Myc* knock-in mouse was made according to standard procedures. Using the approach of Henry et al., 2012, we tagged the C-terminus of mouse SUN1 by attaching two copies of superfolder GFP, a variant of GFP with increased brightness and stability (Pédelacq et al., 2006), and six tandem copies of Myc. We inserted this cassette into a *Rosa26* targeting vector (Soriano, 1999) downstream of a *CAG* promoter and a *loxP-3x polyA-loxP* sequence. The construct was electroporated into 129-derived ES cells, and correctly targeted cells were injected into C57BL/6J blastocysts to screen for chimeras. Chimeric males were bred to C57BL/6J females and intercrossed to obtain homozygotes. *R26-CAG-LSL-Sun1-sfGFP-Myc* mice have been deposited at JAX (Stock 021039). *Camk2a-Cre* (Stock 005359), *PV-Cre* (Stock 008069), *VIP-Cre* (Stock 010908), and *Sox2-Cre* (Stock 008454) mice were obtained from JAX.

For each INTACT experiment, the neocortices of one to two mice were rapidly dissected in ice-cold homogenization buffer (0.25M sucrose, 25mM KCl, 5mM $MgCl_2$, 20mM Tricine-KOH). The tissue was minced with a razor blade and Dounce homogenized using a loose pestle in 5 mL of homogenization buffer supplemented with 1mM DTT, 0.15mM spermine, 0.5mM spermidine, and EDTA-free protease inhibitor (Roche 11 836 170 001). A 5% IGEPAL-630 solution was added to bring the homogenate to 0.3% IGEPAL-630, and the homogenate was further dounced with five strokes of the tight pestle. When purifying RNA, RNasin Plus RNase

Inhibitor (Promega N2611) was added at 60 U/mL. The sample was filtered through a 40 μm strainer (Fisher Scientific 08-771-1), mixed with 5 mL of 50% iodixanol density medium (Sigma D1556), underlayed with a gradient of 30% and 40% iodixanol, and centrifuged at 10,000g for 18 minutes in a swinging bucket centrifuge at 4°C. Nuclei were collected at the 30%-40% interface and pre-cleared by incubating with 20 μL of Protein G Dynabeads (Life Technologies 10003D) for 10 minutes. After removing the beads with a magnet, the mixture was diluted with wash buffer (homogenization buffer plus 0.4% IGEPAL-630) and incubated with 10 μL of 0.2 mg/mL rabbit monoclonal anti-GFP antibody (Life Technologies G10362) or anti-Myc antibody (homemade, ~2 μg) for 30 minutes. 60 μL of Dynabeads were added, and the mixture was incubated for an additional 20 minutes. To increase yield, the bead-nuclei mixture was placed on a magnet for 30 seconds to 1 minute, completely resuspended by inversion, and placed back on the magnet. This was repeated 5-7 times. Bead-bound nuclei were passed through a 20 μm strainer (Partec 04-0042-2315) and washed with 2 x 10 mL, 1 x 2.5 mL, and 1 x 1 mL wash buffer. All steps were performed on ice or in the cold room, and all incubations were carried out using an end-to-end rotator.

All calculations of INTACT specificity and yield used pooled neocortices (approximately dorsal 2/3 of cortex) of two adult mice. To calculate the specificity of mouse INTACT, bead-bound nuclei were stained with DAPI, viewed by fluorescence microscopy, and the numbers of GFP+ and GFP- nuclei were counted (100-200 nuclei per experiment). To calculate the yield of mouse INTACT, input nuclei (i.e., after step 2 in Figure 1C) and bead-bound nuclei were stained with DAPI. The yield was determined from the total number of input nuclei, the % of GFP+ nuclei in the input, and the total number of bead-bound nuclei after INTACT purification (all quantified by fluorescence microscopy or hemocytometer, 100-200 nuclei per experiment). All mice used for INTACT experiments were 8-11 weeks old.

Immunohistochemistry and Microscopy

Mice were anesthesized with ketamine/xylazine, perfused with 4% paraformaldehyde (PFA), and post-fixed for 1 hour at room temperature. Brains were sectioned at 100 µm using a vibratome. Sections were blocked with 10% NGS and 0.25% Triton in PBS and incubated with the following antibodies overnight at 4°C: rabbit anti-Parvalbumin (1:5000, Swant PV 25), rabbit anti-Vasoactive intestinal peptide (1:200, ImmunoStar 20077), mouse anti-NeuN (1:500, Millipore MAB377). Either chicken anti-GFP (1:500, Aves GFP-1020), rabbit anti-GFP (1:400, Life Technologies A11122), or rabbit anti-Myc (1:50,000, homemade) was co-incubated. For staining with anti-VIP, mice were perfused with 2% PFA as heavy fixation decreased the VIP signal. For mouse anti-GAD67 (1:800, Millipore MAB5406), no Triton was included in any step, and both primary and secondary antibody incubations were performed at room temperature for 36 hours. For fluorescent labeling, the sections were incubated with Alexa Fluor-conjugated IgG secondary antibodies (1:400, Life Technologies) and DAPI before mounting with Fluoromount G (SouthernBiotech). Images were taken using a Zeiss LSM700 confocal microscope with minor processing using ImageJ and Adobe Photoshop. For assessment of Cre driver specificity, we counted more than 200 neocortical nuclei for each mouse and two mice per Cre driver.

Flow cytometry

Beads-only control, input nuclei, and bead-bound INTACT-purified nuclei (using anti-Myc antibody) from *VIP-Cre; R26-CAG-LSL-Sun1-sfGFP-Myc* neocortices as well as beads-only control, input nuclei, and bead-bound INTACT-purified nuclei (using anti-GFP antibody) from *PV-Cre; R26-CAG-LSL-Sun1-sfGFP-Myc* neocortices were analyzed using a MoFlo MLS high-speed cell sorter (Beckman Coulter).

Extraction of RNA, DNA, and native nucleosomes

Bead-bound nuclei or whole neocortical nuclei were directly resuspended in Buffer RLT for RNA purification using the RNeasy Micro kit (Qiagen 74004) following the standard protocol with on-column DNase digestion. For RNA preparation from whole neocortical nuclei, nuclei were prepared identically to the INTACT procedure, except that the 40% iodixanol layer was omitted, and nuclei were pelleted by centrifugation and resuspended in Buffer RLT. Bead-bound nuclei were resuspended in PBS for DNA purification using the DNeasy Blood and Tissue kit (Qiagen 69504). Nucleosomes for native ChIP-seq were prepared as previously described (Henry et al., 2012). Briefly, 1-2 million bead-bound nuclei were digested with 0.025 units/µL micrococcal nuclease (Worthington LS004798) in 500 µL of 15mM HEPES pH 7, 1mM KCl, 2mM $MgCl_2$, 2mM $CaCl_2$, 340mM sucrose, 0.15mM spermine, 0.5mM spermidine, and 5mM sodium butyrate at 37°C for 15 minutes. The reaction was terminated by the addition of EGTA to 2mM final concentration. Nucleosomes were extracted for 30 minutes on ice with 200 µL 15mM HEPES pH7, 200mM NaCl, 25mM KCl, 2mM $MgCl_2$, 1mM EGTA, 340mM sucrose, 0.15mM spermidine, 0.15mM spermine, and 5mM sodium butyrate. A second 30 minute extraction was performed with the same buffer except the salt concentration was raised to 400mM NaCl. The extracts were combined and dialyzed overnight against 15mM HEPES pH7, 25mM KCl, 1mM β-mercaptoethanol, 1mM PMSF, and 5mM sodium butyrate using a 10K cut-off Slide-a-Lyzer dialysis device (Thermo Scientific 88401).

RNA-Seq library construction and sequencing

RNA quality was measured by an Agilent Bioanalyzer, with RIN scores consistently greater than 8. Total RNA (2-50 ng) was converted to cDNA and amplified using Nugen Ovation RNA-seq System V2 (Nugen 7102). All RNA samples received a 1:10,000 dilution of ERCC RNA (Life Technologies 4456740). Amplified cDNA was fragmented, end-repaired, linker adapted, and sequenced for 50 cycles on an Illumina HiSeq 2500 instrument. Image analysis and base calling were performed with the standard Illumina pipeline versions RTA 1.12.4.2 and 1.17.20.

MethylC-Seq library construction and sequencing

MethylC-seq libraries were constructed as previously described (Lister et al., 2013), except that samples were PCR amplified with KAPA HiFi HotStart Uracil+ ReadyMix (Kapa Biosystems KK2802) using the following PCR conditions: 2 minutes at 95°C, 4 cycles of [15 seconds at 98°C, 30 seconds at 60°C, 4 minutes at 72°C], and 10 minutes at 72°C. Libraries were sequenced on an Illumina HiSeq 2000 up to 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0.

ATAC-Seq library construction and sequencing

Approximately 50,000 bead-bound nuclei were transposed in a 50 μL volume of 1X TD buffer and 2.5 μL Tn5 transposase (Illumina FC-121-1030) for 30 minutes at 37°C, as previously described (Buenrostro et al., 2013), with the modification that fragmented genomic DNA was recovered using Buffer QG coupled with MinElute spin columns (Qiagen 28604). Transposed genomic DNA was amplified by five cycles of quantitative PCR. 10% of the PCR was subjected to an additional 20 cycles of SYBR green-based qPCR while the remainder of the sample was left on ice. Analysis of the qPCR data allowed a rough estimate of the number of additional cycles needed to generate product at 25% saturation. Typically, four to seven additional PCR cycles were added to the initial set of five cycles. Amplified DNA was purified on AMPure XP beads (Beckman A63881), analyzed on an Agilent Bioanalyzer, and sequenced (paired-end) on an Illumina HiSeq 2500 for 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline versions RTA 1.17.20 and 1.17.21.3.

ChIP-Seq library construction and sequencing

We used the HT ChIP-seq protocol (Garber et al., 2012) for the ChIP reactions and subsequent library construction with the following modifications. For each reaction, chromatin prepared from 0.5-1 million nuclei was incubated with 1 μg antibody and 25 μL Protein G Dynabeads. The following antibodies were used: rabbit anti-H3K27me3 (Millipore 07-449), rabbit anti-H3K27ac (Abcam ab4729), rabbit anti-H3K4me3 (Abcam ab8580), and rabbit anti-H3K4me1 (Abcam ab8895). ChIP-enriched and input DNA was end-repaired, linker adapted, amplified, and sequenced on an Illumina HiSeq 2500 for 50 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 1.17.20.

Fluorescent *in situ* hybridization

cDNA libraries were generated from neocortical 8 week old C57Bl/6J brains with the SuperScript III First-Strand Synthesis System (Life Technologies 18080-051). The following primers were used for producing probes: *3110035E14Rik* (For: 5’-GATAAGAAAGCACTGTGGTCCC-3’, Rev: 5’-ACAGTGAGAAAATCCACCCAAG-3’); *Rasal1* (For: 5’-GTGTGTTCTGGGGCAACC, Rev: 5’-GCTTCTCCACACACCGCT-3’); *Scube1* (For: 5’-TGGACTAGGTGTTGTGTGGAAG-3’, Rev: 5’-TAGCTTCTCCCTGAGTTCCAAG-3’); *6330403A02Rik* (For: 5’-GGCATGCTTATCCAACTACACA-3’, Rev: 5’-TACATTTCATGAGTCCCAGTGC-3’); *Kcng4* (For: 5’-CCATCCCATGGCTGAGAC-3’, Rev: 5’-CAGCATTAGCCCCCATTG-3’); *Afap1* (For: 5’-CAGCAAGGCACAGACCCT-3’, Rev: 5’-TGACTGCTGGGAGCCTTC-3’); *Prss23* (For: 5’-GGGGCAGGATCCACTTCT-3’, Rev: 5’-AGCAGCGTGGGAATTCTG-3’); *Inpp5j* (For: 5’-CTTTCAACTTTGTGCTGGTGAG-3’, Rev: 5’-GTAACCCAGAATGAAGTCTCCG-3’); *9930013L23Rik* (For: 5’-ATCTGGGTGACTCTGGAGAC-3’, Rev: 5’-AGAGGCCACCTCTTCTCTC-3’); *Zfp536* (For: 5’-TATCAGGCCTGGCAGCTC-3’, Rev: 5’-AGTCGATTCCGGGGAGAC-3’); *Slc17a7* (For: 5’-CAGAGCCGGAGGAGATGA-3’ ; Rev: 5’-TTCCCTCAGAAACGCTGG-3’); *Pvalb* (For:

5’-TCTGCTCATCCAAGTTGCAG-3’ ; Rev: 5’-TCCTGAAGGACTCAACCCC-3’); *Vip* (For: 5’-CCTTCCCTAGAGCAGAACTTCAG-3’ ; Rev: 5’-ACATCAATTTTCCTCGATTGCTAC-3’). For all genes except *9930013L23Rik*, we used the same primers as the Allen Brain Atlas (http://mouse.brain-map.org/) (Lein et al., 2007). Standard methods for dual color fluorescent *in situ* hybridization were used. Briefly, adult C57Bl/6J brains were fresh-frozen in OCT compound and 20 µm sections were cut. After probe hybridization and post-hybridization washes, the sections were incubated with 3% hydrogen peroxide in PBS to quench endogenous peroxidase activity. The DIG-labeled probe (candidate cell type-enriched gene) was detected with anti-DIG-POD (Roche 11207733910) followed by TSA Plus amplification (Perkin Elmer NEL745001KT). After quenching with hydrogen peroxide, the fluorescein-labeled probe (*Slc17a7*, *Pvalb*, or *Vip*) was detected with anti-Fluorescein-POD (Roche 11426346910) followed by TSA Plus amplification (Perkin Elmer NEL741001KT).

B. General Data Analysis

Data processing steps made extensive use of Bowtie (Langmead et al., 2009; Langmead and Salzberg, 2012), Tophat (Trapnell et al., 2009), BEDTools (Quinlan and Hall, 2010), and custom scripts. All reads were aligned to the mm10 genome. Browser representations were created using AnnoJ (http://www.annoj.org) (Lister et al., 2009). Correlations are Pearson, unless otherwise indicated.

RNA-seq data processing

We aligned reads from the libraries in two ways: 1) to the whole genome for genome browser visualization and 2) to the annotated transcriptome to estimate gene expression levels. RNA-seq reads were trimmed (*seqtk trimfq* -b 5) before aligning to the genome (TOPHAT v1.4.0) for visualization. Gene expression levels were estimated using RSEM v1.1.20 (Li et al., 2011) calling BOWTIE v0.12.7 for protein-coding genes (mm10 iGenomes annotation). Differentially expressed genes (5% FDR) were identified through pairwise comparisons using EBSeq (v1.1) (Leng et al., 2013). Additional RNA-seq data measured from fetal cortex and 6 week cortex (Lister et al., 2013) were also processed with RSEM. We used TF annotations from AnimalTFDB (Zhang et al., 2012) for all analyses focused on TFs.

MethylC-seq data processing

MethylC-seq reads were processed as previously described using the *methylpy* pipeline (https://bitbucket.org/schultzmattd/methylpy/ and Lister et al., 2013). Briefly, all cytosines in the forward and reverse complement strands of the mm10 reference genome (appended with the lambda phage genomic sequence) were converted to thymines followed by bowtie index building using the *build_ref* function. The mapping of MethylC-seq reads was performed with the *run_methylation_pipeline* function. Adapters in MethylC-seq reads were trimmed using *cutadapt*. All cytosines in the trimmed reads were then computationally converted to thymines and mapped to a converted forward strand reference and to a converted reverse complement strand reference. Reads were only allowed to map to one location, and clonal reads were removed. The resulting datasets were stored as “allc” tables containing one row for each genomic cytosine position and columns representing the genomic context (e.g. CG, CH); the number of reads supporting a methylcytosine at that position (mc); and the total number of reads at that position (h).

Many of our analyses are based on profiling the methylation level in CG and CH contexts (i.e., % mCG and % mCH) at a particular site, within a genomic region, or in a set of regions. The methylation level at a set of positions R is defined as:

$$[\%\, mC]_R = 100 * \sum_{i \in R} mc_i \Big/ \sum_{i \in R} h_i$$

For some analyses, we adjusted these estimates to correct for bisulfite non-conversion (see below, methods for Figure 3C).

DMR Finding

We estimated DMRs using a previously reported method (Ma et al., 2014; Schmitz et al., 2013), which is implemented in the *DMRfind* function of *methylpy* (available at https://bitbucket.org/schultzmattd/methylpy/). Since we observed a high degree of consistency between biological replicates, we pooled reads from replicates for DMR calling to enhance the statistical power. DMR calling used five samples: E13 fetal cortex and S100b+ sorted glia (Lister et al, 2013) and pooled replicates of excitatory, PV, and VIP neurons. After identifying individual CG cytosines with differential methylation, blocks that contained fewer than 4 differentially methylated sites were discarded.

DMRs and mice strain differences

Our identification of DMRs does not factor in SNPs or indels across mice strains. The presence of strain-specific genetic variants could potentially affect our estimates of methylation levels from MethylC-seq data. This could affect the identification of DMRs as the INTACT mice used in this study are from different genetic backgrounds. In spite of this, genome-wide DNA methylation data from excitatory neurons highly correlated with NeuN+ data from inbred C57Bl/6J mice. Furthermore, we saw a high correlation between the excitatory neuron methylome and the NeuN+ methylome at localized differentially methylated regions.

To address how SNPs and indels could affect the identification of differentially methylated regions, we obtained SNPs and indels (relative to the reference C57BL/6J genome) for three mice strains whose genomes are available (129S1, 129P2, C57Bl/6N; http://www.sanger.ac.uk/resources/mouse/genomes/). *R26-CAG-LSL-Sun1-sfGFP-myc* mice are on a mixed 129;C57BL/6J genetic background which includes 129S1. *PV-Cre* mice are on a mixed 129P2;C57BL/6J background. *Camk2a-Cre* mice are on a mixed C57BL/6J;C57BL/6N background. Although we do not have a similar SNP or indel list for the 129S4Sv/Jae strain (present in *VIP-Cre*) or 129X1 (also present in *R26-CAG-LSL-Sun1-sfGFP-myc*), we expect that the majority will be present in one of the other 129 strains (129S1 and 129P2). In support of this, 72.2% of all the indels and 87.5% of the SNPs that appear in either 129 strain are common to both of them (867,258/1,200,566 of indels and 4,988,081/5,697,417 SNPs).

First, by plotting the density of SNPs and indels relative to DMR locations, we found a small depletion of strain-specific variants around DMRs (data not shown), suggesting that our DMR calling is not inflated by the presence of strain-specific variants. Next, we examined 419,626 SNPs and indels that overlapped a CG position, as the DMR finder only evaluates CG sites. We then re-ran the DMR caller after removing the overlapping CG sites. 245,383 DMRs were identified using this masked data. Of these, 99.99% (245,354) overlap with the original 251,301 DMRs. Out of the original 251,301 DMRs, 97.6% (245,266) overlap with the masked DMRs. Based on this analysis, we used the original set of 251,301 DMRs in the manuscript.

It remains possible that some strain-specific genetic variants may directly affect methylation levels. In those cases it could be the case that differential methylation is driven by strain genotype differences rather than cell type differences. Although we cannot rule this out, we think the vast majority of DMRs are cell type-driven rather than strain-driven, for the following reasons: (1) INTACT-purified excitatory neuron and NeuN+ methylomes are extremely consistent; (2) The animals are all mixed backgrounds, so strain-derived genetic components segregate independently. The high correlation between replicates argues against substantial variant effects contributing to DMR calling; and (3) The consistency of our results with known cell type markers.

Comparison of DMR finding using NeuN+ versus INTACT-purified excitatory neuronal nuclei

To compare the number of total and cell type-specific DMRs identified using INTACT-purified excitatory neurons versus NeuN+ neurons from Lister et al., 2013, the MethylC-seq data for NeuN+ nuclei from the 7 week-old male (SRX314951) and the 12 month-old female (SRX314955) were combined to best match the coverage of the excitatory neuron methylomes. Identical DMR calling procedures were performed except that the NeuN+ sample was substituted for the excitatory neuron sample.

Identifications of UMRs, LMRs, and large DNA methylation features

UMRs and LMRs were identified using MethylSeekR (Burger et al., 2013) with m = 0.5 and 5% FDR. MethylSeekR did not classify any regions as partially methylated domains. DMVs were identified as UMRs ≥5 kb with mean.meth (column 7 in MethylSeekR output) ≤15. To identify large hypo-DMRs, all hypo-DMRs for each cell type with inter-DMR distances less than 1 kb were merged (*bedtools merge* -d 1000). Merged hypo-DMRs were further stratified into those ≥2 kb (called "large hypo-DMRs") and those <2 kb. mm10 CpG island annotations were downloaded from the UCSC table browser.

Estimation of hmC at DMVs

To estimate the contribution of hmC to the excitatory neuron hyper-methylation of DMVs associated with *Neurog2* and *Pax6*, we mapped 6 wk cortex TAB-seq data from Lister et al., 2013 to mm10, calculated the % hmC in the region defined by Figure 13E, performed correction for non-conversion and protection, and compared it with the MethylC-seq signal of excitatory neurons in the same region.

ATAC-seq data processing

Adapter sequences were trimmed from ATAC-seq reads (*cutadapt* v1.3 -a CTGTCTCTTATACACATCT -q 30 --minimum-length 36 --paired-output), before aligning (BOWTIE2 v2.1.0 -t -X2000 --no-mixed --no-discordant) and removing redundant reads (*picard MarkDuplicates*). Fragment ends were offset by 4nt towards the center of each fragment.

Peaks were called with HOMER (*findPeaks* -region -size 500 -minDist 50 -o auto -tbp 0) (Heinz et al., 2010) using sub-nucleosomal (<100 bp) fragments, and overlapping peaks were merged (*bedtools merge*). Peaks called from replicates were merged (*bedtools merge*) to yield a peak set for each cell type. We used *bedtools multiinter* to classify peaks as cell type-specific or shared, keeping only those regions greater than 100 bp. For analyses of cell type-specific versus shared peaks, the peaks following *bedtools multiinter* were used.

Footprinted sites were predicted using CENTIPEDE on ATAC-seq fragments of all lengths (Pique-Regi et al., 2011). TF binding matrices were obtained from the MEME motif database (v11, 2014 Jan 23. motif sets chen2008, hallikas2006, homeodomain, JASPAR_CORE_2014_vertebrates, jolma2010, jolma2013, macisaac_theme.v1, uniprobe_mouse, wei2010_mouse_mw, wei2010_mouse_pbm, zhao2011) and scanned across the mouse genome to identify hits using FIMO (--output-pthresh 1E-5 --max-stored-scores 500000) (Bailey et al., 2009; Grant et al., 2010). For every ATAC-seq sample, we counted the frequency of Tn5 insertion events in 200 bp windows centered at every motif instance in the genome using *bwtool* (Pohl and Beato, 2014). These count matrices were then used by CENTIPEDE along with conservation levels at corresponding positions (phyloP score from the placental subset of the UCSC 60-way genome alignment; Karolchik et al., 2014) to learn motif-specific models of Tn5 insertion density and predict the likelihood that each motif instance across the genome is bound. We used sites predicted with greater than 95% posterior probability to be occupied as our footprint set.

To predict nucleosome positions, we employed the same procedure as Buenrostro et al., 2013. First, an estimated set of mononucleosomal fragments was generated by classifying fragments into sub-, mono-, di-, tri-, tetra-, and penta-nucleosomal fragments using a mixture of gaussians fitted to the length distribution from each sample (*mixtools* package in R). Multi-nucleosomal fragments were split into single nucleosomes by fragmenting them uniformly into the number of nucleosomes they were predicted to span, only considering those fragments whose numbers of nucleosomes were predicted with greater than 90% posterior probability. We then estimated nucleosome positioning by subtracting the "background" signal of sub-nucleosomal fragments from the "foreground" of mono-nucleosomal fragments (DANPOS -x 1 -k 1 -p 1 -a 1 -d 20 --clonalcut 0) (Chen et al., 2013).

ChIP-seq data processing

Excitatory neuron histone modification ChIP-seq and input reads were aligned (BOWTIE v0.12.7 -m 1), and redundant reads were removed (*samtools rmdup*). CREB, SRF, and NPAS4 ChIP-seq and input reads (Kim et al., 2010) generated by SOLiD^{TM} sequencing were aligned using BOWTIE v1.0.0 in colorspace mode (-C). FOS, FOSB, and JUNB ChIP-seq and input reads (Malik et al., 2014) were aligned using BOWTIE2 2.0.2 followed by the removal of reads with mapping quality score below 20.

We used SICER (Zang et al., 2009) to identify H3K4me1, H3K4me3, H3K27ac, and H3K27me3 ChIP-seq peaks from excitatory neurons. SICER (SICER_V1.1) parameters with FDR = 0.001 were: redundancy threshold=1; fragment size=150; W=200, G=200 for H3K4me1, H3K4me3, and H3K27ac; and W=200, G=1000 for H3K27me3.

TF ChIP-seq peaks were identified using MACS 1.4 (Zhang et al., 2008) with a p-value cutoff at 1E-10.

Identification of putative enhancers in excitatory neurons

Putative enhancers were defined by combining H3K27ac and H3K4me1 SICER peaks. Regions overlapping with H3K4me3 peaks or ± 2.5 kb from an annotated TSS were removed to exclude promoter features.

DNaseI-seq data processing

We obtained DNaseI-seq data from 53 samples across a diverse set of neuronal and non-neuronal tissues from the mouse ENCODE project (Stamatoyannopoulos et al., 2012). Datasets were processed using the ATAC-seq analysis pipeline modified for single-end reads. Only uniquely aligning reads were kept (BOWTIE v0.12.7, options -m 1). We identified peaks of DNaseI sensitivity using HOMER (*findPeaks* -region -size 500 -mindist 50 -o auto -tbp 0). To compare DNaseI-seq peaks from whole cerebrum (GSM1014168) versus ATAC-seq peaks, we calculated the percentage of ATAC-seq peaks overlapping DNaseI-seq peaks (union of peaks in three replicates).

C. Figure-Specific Data Analysis

Figure 4B (comparison of INTACT and manually sorted RNA in PV neurons)

Microarray data from manually sorted GFP+ neurons in P40 G42 transgenic mice (downloaded from GSE17806; Okaty et al., 2009) was processed using *rma* normalization from the R package *affy*. To aid in visualization, *rma*-normalized microarray values were transformed back to the linear scale, before plotting both RNA-seq TPM values and microarray values on a log scale. The Spearman correlation coefficient was calculated in R (*cor* with the option "spearman").

Figure 3C (global DNA methylation level)

We adjusted the methylation level for the effect of bisulfite non-conversion, which was calibrated in each experiment by sequencing of spiked-in unmethylated phage-lambda DNA. The non-conversion rate, s, ranged from 0.29% to 0.38% across our samples. We used the maximum likelihood estimate for the true methylation level (% mC) by adjusting for non-conversion as follows:

$$[\%\ mC]_{max.likelihood} = 100 * G[\frac{mc/h - s}{1 - s}]$$

where mc, h are the number of methylated cytosine base calls and the total cytosine base calls within a region, respectively, and $G[x] \equiv min[max[x, 0],1]$ ensures that the estimated methylation level is in the interval [0,1].

Figures 3E and 4E (line plots of % mCG and % mCH in highly cell type-specific genes)

To assess the pattern of methylation around specific groups of DE genes, we first pooled methylation data from biological replicates. For each gene, we profiled % mCG and % mCH within 1 kb bins between 100 kb upstream of the transcription start site (TSS) and the TSS and between the transcription end site (TES) and 100 kb downstream. We divided each gene body into 10 equally spaced bins. When multiple transcripts shared the same TSS and TES, we only used 1 instance for the analysis. We then computed the % mC (corrected for bisulfite non-conversion) for each bin. Gene lists were filtered by requiring >5 fold-change and ≥0.95 posterior probability of differential expression (PPDE) from EBSeq. We focused on genes that are differentially over-expressed in one cell type relative to both of the other cell types. These DE genes were also used for Figures 9E-F.

Figure 4I (comparison of intragenic and genomic % mCH across replicates and cell types)

For each 5 kb genomic bin or gene body, we computed the CH methylation levels in each sample and corrected for bisulfite non-conversion. We excluded any genomic bins or genes with low coverage (<50 base calls) or genes with short (<500 bp TSS-TES) length. We normalized the mCH level for each sample by the median across all genomic bins or gene bodies. We then computed the ratio of the methylation levels between cell types (solid lines). As a control, we also computed this ratio for comparisons of biological replicates (dashed lines). In comparisons where both samples had very low levels of mCH (<0.5%), we set the fold-change to 1.

Figures 5B and 5D (Venn diagrams of ATAC-seq peaks and DMRs)

Venn diagrams were created using euler*APE* (Micallef and Rodgers, 2014).

Figure 6B, right (activity-dependent TF peaks at hypo-DMRs)

For TF *i*, the enrichment or depletion of each hypo-DMR category *j* overlapping each TF *i* ChIP-seq peak category, relative to all DMRs, was represented as: $\log_2$(fraction of category *j* hypo-DMRs overlapping TF *i* peaks / fraction of all DMRs overlapping TF *i* peaks). The hypergeometric test (MATLAB *hygecdf*) was used to test for significance: *hygecdf(number of category j hypo-DMRs associated with TF i ChIP-seq peaks, sample size of all DMRs, number of all DMRs associated with TF i ChIP-seq peaks, sample size of category j hypo-DMRs)*. The option ‘upper’ was applied for testing enrichment.

Figures 5E and 6C (levels of CG and CH DNA methylation, ATAC-seq reads, histone ChIP-seq reads, and activity-dependent TF ChIP-seq reads at DMRs)

Hypo-methylated cell types in DMRs were identified from *methylpy*, and only DMRs with one or two hypo-methylated cell types were displayed in the heatmap. Wiggle files for DNA methylation levels were created from *methylpy run_methylation_pipeline* output files at 100 bp resolution for CG and CH contexts. Wiggle files for ATAC-seq were generated by the pileup of sub-nucleosomal (<100 bp) reads at 100 bp resolution. Wiggle files for histone modification and TF ChIP-seq were generated using MACS14 with options -w and --space 100. Profiles of DNA methylation, ATAC-seq (normalized for library size), and ChIP-seq (normalized for library size) were plotted in a 3 kb region centered at DMRs using the wiggle files.

Figure 6D (correlation between CG and CH methylation at DMRs)

% mCG and % mCH levels for excitatory, PV, and VIP neurons in each DMR were normalized by the mean % mCG and % mCH level across the three cell types for that DMR. The Pearson correlation between normalized mCG and mCH was calculated with the MATLAB *corr* function.

Figure 7B (correlation across cell types of ATAC-seq density at peaks)

The similarity of ATAC-seq read distributions between pairs of ATAC-seq samples was quantified using the Pearson correlation of read densities over the union of peaks called across all samples (*deeptools bamCorrelate*) (Ramírez et al., 2014).

Figure 10A (evaluation of DNA methylation at footprints)

To investigate the correlation between TF binding and local DNA methylation, we focused on footprints that were unique to one of the three cell types; that is, we excluded footprints that had the same start and end site in more than one cell type. Then, for each footprint of a given TF, we calculated methylation levels (% mCG, % mCH, and % mCA) within ±50 bp of the footprint start site; this value is the methylation level for “FP present” locations (y-axis). We compared this with the methylation levels at the same location in the other cell types (“FP absent,” x-axis), provided that the TF is expressed (TPM≥30) in these other cells. Then, the average methylation level using all MethylC-seq reads from footprinted regions (i.e., excitatory reads at excitatory-specific FPs, PV reads at PV-specific FPs, and VIP reads at VIP-specific FPs) was plotted against the average methylation level of all MethylC-seq reads from non-footprinted regions (i.e., excitatory and PV reads at VIP-specific FPs, etc.).

Figures 9B and 10B (enrichment of TF footprints and hypo-DMR motifs)

For Figure 9B, we ranked each TF by the relative enrichment of their footprints in a foreground category of cell type-specific ATAC-seq peaks versus a background of the other two categories of cell type-specific ATAC-seq peaks (e.g., footprints in excitatory-only peaks versus PV-only and VIP-only peaks) and additionally required that the TF itself be expressed (≥30 TPM) in the foreground cell type (e.g., in excitatory neurons). The significance of enrichment was estimated using the pairwise Fisher's test (*pairwise.Fisher.test* in the *fsmb* R package) (i.e., to compare the ratio of a TF's footprints to the total number of footprints predicted in one cell type against the corresponding ratio computed from footprints in the other two cell types). The same test was used to compare the enrichment of a TF's motifs in one category of cell type-specific hypo-DMRs versus the other two categories of cell type-specific hypo-DMRs.

To assess TF motifs that were enriched in DMRs hypo-methylated in one or two INTACT-purified cell types as well as fetal cortex and glia (Figure 10B), hypergeometric tests were performed for each TF motif using the occurrence of the motif in all DMRs as the background.

Figures 9D (construction of TF-TF regulatory networks)

TF A was predicted to regulate TF B when: (1) TF A was expressed in a cell type (≥30 TPM), (2) TF A had a predicted footprint in a cell type-specific ATAC-seq peak, (3) the ATAC-seq peak was within 10 kb of the TSS for TF B, and (4) TF B was expressed in that cell type (≥30

TPM). The resulting set of predicted regulatory interactions was visualized as a network (*igraph* package in R), omitting TFs with more than 20 connections to ease visualization. To define a pan-neuronal regulatory network, we identified footprints common to all three cell types that occurred in shared ATAC-seq peaks and did not overlap ubiquitous DNaseI peaks (peaks occurring in at least 40 out of 53 processed DNaseI-seq samples).

Figures 12A-B (sparse generalized linear model of mRNA expression)

To assess how well mRNA expression levels correlate with a combination of epigenetic and chromatin features, we fit a generalized linear model using the MATLAB implementation of *cvglmnet* (Friedman et al., 2010) with the Poisson distribution and parameter $\alpha = 1$ (corresponding to LASSO regression). This model assumes Poisson distributed noise and uses LASSO regularization to promote sparseness, i.e. to fit the model using a small subset of features. We used 10-fold cross-validation to avoid overfitting and default values for all other parameters. For each gene, we used the longest isoform to guarantee that each gene contributes only once to the dataset and there is no overlap between training and test sets. To define features for this analysis, we created 18 windows of varying sizes, ranging from 200 bp to 32 kb, surrounding each TSS. For mCG and mCH we also included two additional windows for the gene body and the flanking region. Within each window we computed the value of mCG, mCH, ATAC-seq, and DMR density, resulting in a total of $4 \times 18 + 2 \times 2 = 76$ features. 7 parameters in the regression model achieves 1 standard error above the minimum cross-validated error.

Figures 11B-E and 12C (k-means clustering of genes by intragenic mCH followed by assessment of gene expression and ATAC-seq peak enrichment in each cluster)

To identify sets of genes that share similar DNA methylation patterns in an unbiased fashion, we applied k-means clustering to the gene body mCH. We profiled % mCH in gene bodies (TSS-TES) within each of eight samples included in this analysis (Fetal and Adult Cortex, NeuN+, NeuN-, and Glia from Lister et al., 2013; Exc, PV, and VIP from the current study). We excluded 468 genes with short gene bodies (<500 bp TSS-TES) or with low coverage in our methylome datasets (<50 cytosine basecalls within the gene body in any sample); the remaining 23,023 genes were included. When multiple transcripts shared the same TSS and TES, we only used 1 instance for the analysis. To compensate for the differing genome-wide background level of % mCH in different cell types, we normalized the methylation level in each sample for each gene by the average over the gene's distal flanking region (50-100 kb upstream of TSS or downstream of TES). We then log-transformed the normalized methylation levels. Next, we used the MATLAB function *kmeans* to apply k-means clustering using data from five samples representing distinct cell types or developmental stages (Fetal cortex, Exc, PV, VIP, and Glia). Clustering used 1 minus the correlation coefficient of normalized mCH values across genes as a distance measure. We chose to extract *k*=25 clusters to capture a diverse range of methylation features, while still allowing visualization and statistical enrichment analysis of functional association for each gene set (Figure 12C). We repeated the clustering procedure five times using random initialization of the cluster centers, choosing as the final estimate the run with the smallest within-cluster sum of distances from each point to the cluster centroid.

To display the CH methylation patterns within these gene clusters in Figure 11B, we profiled % mCH in 1 kb bins starting 100 kb upstream of the TSS and ending 100 kb downstream of the TES. To compare genes with different lengths, we divided each gene body into 10 non-overlapping bins of equal size extending from the TSS to the TES. Methylation levels were normalized by the flanking region as described above. We then linearly interpolated the gene-body mCH data at 100 evenly spaced bins within the gene body in order to give similar visual weight to the gene-body and flanking methylation data. Finally, we smoothed and downsampled the genes 40-fold to allow representation of genome-scale features.

RNA TPM levels were plotted for each cluster (Figure 11C). Last, we assessed the enrichment of specific categories of DE genes (Figure 11D) and ATAC-seq peaks located within ±10 kb of each gene's TSS (Figure 11E) within each cluster using hypergeometric test with Benjamini-Hochberg FDR control.

Figures 11F and 12D, top (DNA methylation levels relative to nucleosome calls)

For each nucleosome call, we counted the number of sequenced CG (Figure 12D, top) or CH (Figure 11F) base calls starting from the nucleosome center up to 2000 bp away. We also counted the number of these sequenced base calls that were methylated. The ratio of the methylated to the total number of sequenced base calls is the average mCG or mCH level at that position. Because we were able to average over all of the estimated nucleosome positions, binning was not necessary and the mCH level was estimated as a function of distance with 1 bp resolution. We determined the average in the flanking region by summing over all base calls from 1-2 kb upstream and downstream of each nucleosome call. This average is a single number, not a function of distance from the nucleosome. The normalized curves in the figures show the mCG and mCH level divided by the flanking region average.

Figures 14C-D (GC content and CG methylation level of hypo-methylated features)

GC contents were computed with *bedtools nuc* (Figure 14C). Genomic regions matching the sizes of excitatory hypo-DMRs < 2 kb were randomly selected with *bedtools shuffle*. The random selection was repeated 100 times.

Figure 13C (enrichment of histone marks, ATAC-seq reads, and RNA over hypo-methylated features)

For excitatory neurons, we divided DMVs into those that overlapped with SICER peaks for H3K4me3 by ≥1 bp or those that overlapped with SICER peaks for H3K27me3 by ≥1 bp. For each type of DNA methylation feature in excitatory neurons, the $\log_2$ enrichment of each histone modification over the input (both normalized for library size) was plotted (left). For ATAC-seq, $\log_2$(1+ATAC-seq <100 bp reads pileup per 10 million reads) was plotted. Protein-coding genes were associated with all hypo-methylated features if ≥1 bp overlap was found between these features and the region from 10 kb upstream of the TSS to the TES. Log_2(TPM+1) values were plotted for all associated genes.

Figure 13D (overlap of DE genes in hypo-methylated regions)

Differentially expressed (DE) genes identified by EBSeq with ≥2 fold-change were used for this analysis. Protein-coding genes were associated with all hypo-methylated features identically to Figure 13C. The significance of the overlap between each DNA methylation feature and DE genes was tested by hypergeometric distribution using MATLAB *hygecdf* function with *hygecdf(number of category i feature overlapping with category j DE genes, sample size of all DNA methylation features, number of all DNA methylation features that overlap with category j DE genes, sample size of category i feature).* The option 'upper' was applied for testing enrichment. *All DNA methylation features* were defined as combined large hypo-DMRs, merged hypo-DMRs with length less than 2 kb, and DMVs identified for Exc, PV, and VIP neurons. Significance was set at q<1E-5.

Figure 14G (gene ontology enrichment in DMVs)

GO enrichment for each group of DMVs was performed using GREAT (McLean et al., 2010). For the background, DMVs were combined with UMRs between 1-3.5 kb and mean mCG ≤15%.

Chapter III

Results

A. Mammalian INTACT Isolates Specific Populations of Neuronal Nuclei from the Brain

To generate a mouse line for affinity purification of labeled nuclei, we tagged the C-terminus of mouse SUN1, a nuclear membrane protein, with two tandem copies of superfolder GFP and six copies of the Myc epitope (SUN1-sfGFP-Myc). We targeted *Sun1-sfGFP-myc* to the ubiquitously expressed *Rosa26* locus preceded by a *CAG* promoter and a *loxP-3x polyA-loxP* transcriptional roadblock (*R26-CAG-LSL-Sun1-sfGFP-myc*) (Figure 1A). Cells expressing Cre recombinase remove the roadblock and allow transcription of *Sun1-sfGFP-myc*. We first recombined *R26-CAG-LSL-Sun1-sfGFP-myc* in all cells using *Sox2-Cre*, a germline Cre driver (Figure 2A). *Sox2-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice are healthy and fertile with no obvious phenotypic deficits, indicating that long-term expression of the fusion protein is well-tolerated.

We expressed *Sun1-sfGFP-myc* in excitatory (Exc) neurons (*Camk2a-Cre*), PV interneurons (*PV-Cre*), and VIP interneurons (*VIP-Cre*) (Figure 1B). Immunohistochemistry targeting GFP showed that the SUN1 fusion protein is localized to the nuclear periphery. Quantification of labeled nuclei together with neuronal markers (Figures 1B and 2B-G) indicated that each Cre driver predominantly recombines the targeted cell type. The pattern of labeling using anti-Myc is identical to anti-GFP (Figure 2H).

We next developed an affinity purification method to capture GFP+/Myc+ nuclei from fresh tissue homogenates (Figure 1C). Either anti-GFP or anti-Myc antibodies, together with Protein G-coated magnetic beads, can be used to isolate nuclei from both rare and common cell types with high yield and specificity. Examination of input versus affinity purified (anti-GFP) nuclei (Figure 1D) by fluorescence microscopy showed that INTACT achieves >98% purity with >50% yield, even for cell types that represent only 1-3% of the starting tissue (Figure 1E). Similar results were obtained using anti-Myc (95-98% purity with 42-65% yield, n=3). To further assess the specificity of mouse INTACT, we performed flow cytometry on input and affinity purified (anti-Myc) nuclei from *VIP-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice (Figure 2I). Flow cytometry showed that more than 99% of input nuclei (after step 2 in Figure 1C) were singlets, corresponding to well-isolated nuclei, and 1.5% of input nuclei were GFP+. In contrast, 98.9% of affinity purified nuclei were GFP+. Similar results were obtained using anti-GFP (Figure 2J). Therefore, both manual quantification and flow cytometry indicate that mouse INTACT isolates highly pure preparations of tagged nuclei.

B. **INTACT RNA-Seq Captures Neuronal Subtype Markers**

To assess patterns of gene expression and DNA methylation in distinct neuronal subtypes, we used RNA-seq to profile transcript abundance from INTACT-purified nuclei in adult mice, and we used MethylC-seq to generate single base-resolution methylome maps (Lister et al., 2008) from the same cell types, with the caveat that bisulfite sequencing does not differentiate between methylcytosine (mC) and hydroxymethylcytosine (hmC) (Figure 3A). RNA-seq profiles are highly similar across replicates (r=0.98) (Figures 3B, right panel, and 4A). A total of 4,095 genes show ≥2-fold differential transcript abundance across neuronal subtypes, with over 2,000 between each pair of neurons. Established subtype markers are enriched in purified nuclei (e.g., *Slc17a7* and *Dkk3* in excitatory; *Pvalb* and *Lhx6* in PV; *Vip* and *Htr3a* in VIP) whereas markers of other lineages are depleted (Figure 3B, left three panels). The gene expression profile of INTACT-purified PV neurons is also consistent with RNA microarray data from manually sorted PV neurons (Figure 4B). We further used double fluorescent *in situ* hybridization to examine ten genes with previously unknown specificity in neocortical excitatory or PV neurons. Probe labeling for nine out of ten genes co-localized with the neuron type as predicted by RNA-seq and was excluded from other classes (Figure 4C), indicating that INTACT RNA-seq profiles identify novel patterns of gene expression.

C. Non-CG Methylation is a Common Feature of Both Excitatory and Inhibitory Neurons, but Shows Widespread Differences in Genomic Distribution

In our MethylC-seq data, we observed substantial levels of DNA methylation in the non-CG context for all three neuronal populations (Figures 3A, C-D). In most differentiated mammalian cells, DNA methylation is largely confined to the CG dinucleotide context. On the other hand, non-CG methylation (mCH, where H=A, C, or T) is a special feature of adult neurons but accumulates at much lower levels in adult glia and non-neuronal tissues (Lister et al., 2013; Xie et al., 2012). We find that mCH is most abundant in PV neurons (Figure 3C), where it constitutes nearly half (46-47%) of the total methylcytosines (Figure 3D). Because mCH accumulates during the first weeks of post-natal development, coincident with the period of rapid synaptogenesis and long after excitatory and inhibitory lineages have diverged (Guo et al., 2014; Lister et al., 2013), these data suggest that a high level of non-CG methylation is a shared distinction of mature cortical neurons. Furthermore, because all three neuron subtypes share similar motif preferences for mCH, with CAC showing the highest methylation level (Figure 4D), it is likely that a common enzymatic mechanism (Gabel et al., 2015; Guo et al., 2014) is responsible for mCH deposition and maintenance in these neurons.

Both promoter and intragenic DNA methylation in CG and CH contexts inversely correlate with gene expression in the mammalian brain (Lister et al., 2013; Xie et al., 2012).

However, a lack of cell type-specificity in existing *in vivo* datasets can complicate the interpretation at individual genes. For example, *Slc6a1* (GAT-1, primarily expressed in inhibitory neurons) and *Lhx6* (a PV-specific TF) appear to be both actively transcribed and highly methylated in samples of whole cortical tissue and in mixed neurons (NeuN+) (Figure 3A). Our datasets from INTACT-purified nuclei resolve these conflicting signals by showing that active gene expression and DNA methylation do not occur in the same cells, but rather in distinct subpopulations. Using a list of highly specific genes from our RNA-seq data, we find that both intragenic and promoter levels of CH (Figure 3E) and CG (Figure 4E) methylation are higher in the non-expressing cell type.

DNA methylation levels in gene bodies are highly variable across neuronal subtypes. As measured by pairwise Pearson correlations (Figures 3B, F and 4A, F-H), gene body mCH levels are more divergent ($r=0.83\text{-}0.86$) than both gene expression levels ($r=0.95\text{-}0.96$, $p=0.003$, t-test) and mCG levels ($r=0.93\text{-}0.94$, $p=0.001$), whereas biological replicate signals are nearly identical for all features ($r \geq 0.97$). After normalization to adjust for the genome-wide average level of mCH, 8,662 genes (38%) show >50% difference in intragenic mCH in at least one pairwise comparison of cell types, versus 6.1% between biological replicates (Figure 4I, top). Certain genes display notably higher differences. For example, the VIP-specific TF *Prox1* has 23-fold higher mCH in excitatory neurons and 32-fold higher mCH in PV neurons compared to VIP neurons (Figure 3F). Variability in gene body CH methylation is paralleled by extensive differences at a global scale (Figure 4I, bottom). Genome-wide, 37% of all 5 kb bins show >50% difference in mCH between at least one pair of cell types, compared to only 3.8% between biological replicates.

D. **Neuronal Regulatory DNA is Predominantly Cell Type-Specific**

Localized regions of accessible chromatin and low levels of DNA methylation are well-established signatures of *cis*-regulatory elements such as promoters and enhancers (Neph et al., 2012; Stadler et al., 2011; Thurman et al., 2012). Therefore, we mapped the locations of putative gene regulatory regions in specific neuronal subtypes by systematically identifying these two features (Figures 5A and 6A). In excitatory neurons, we also profiled histone modifications using chromatin immunoprecipitation (ChIP) followed by sequencing to identify potential promoters (marked by H3K4me3), enhancers (H3K4me1 and H3K27ac), and Polycomb-associated repressed regions (H3K27me3).

We identified 322,452 discrete peaks of chromatin accessibility (median length 501 bp) in excitatory, PV, and VIP neurons using sub-nucleosomal (<100 bp) reads resulting from *in vitro* transposition of native chromatin by Tn5 transposase (ATAC-seq, Buenrostro et al., 2013). We find that most regulatory elements in neuronal cells are cell type-specific, including the large majority of distal regulatory elements (Figure 5B). In total, only 13.4% (43,354) of ATAC-seq peaks are shared across all three neuronal subtypes. Compared to DNaseI-seq data from the whole cerebrum (Stamatoyannopoulos et al., 2012), nearly all (93%) shared ATAC-seq peaks are also detected as cerebrum DNaseI-seq peaks (Figure 5C). In striking contrast, 62% of VIP-specific, 52% of PV-specific, and 31% of excitatory-specific ATAC-seq peaks are missed in the DNaseI-seq data, highlighting the advantage of INTACT profiling over whole tissue analysis for identifying regulatory regions, particularly those unique to sparse cell types, and for understanding which regulatory regions are active in individual cell types.

We next determined regions that differ in their levels of CG methylation across five cell populations: INTACT-purified excitatory, PV, and VIP neurons, plus fetal embryonic day (E)13 frontal cortex and adult S100b+ glia from Lister et al., 2013. We expected that including purified neurons would facilitate identification of differentially methylated regions (DMRs). Using a conservative statistical approach (Lister et al., 2013), we identified 251,301 DMRs with a median length of 275 bp. 112,462 of these DMRs are hypo-methylated (hypo-DMRs) in excitatory neurons. In keeping with our expectation, substitution of a mixed neuronal sample (NeuN+) with comparable sequencing coverage for the excitatory neuron sample results in 77,417 (68.8%) hypo-DMRs in NeuN+ neurons, despite the prevalence of excitatory neurons in this sample. The increased detection of DMRs using INTACT-purified excitatory neurons again demonstrates the power of cell type-specific profiling for comprehensive identification of regulatory regions. To identify hypo-methylated regions that may not be differentially methylated across cell types, we segmented each methylome into unmethylated regions (UMRs) and low-methylated regions (LMRs) (Burger et al., 2013).

As expected from previous studies (Stadler et al., 2011), the majority of UMRs are located at promoters (66.3-74.2% within 2.5 kb of a TSS) whereas most LMRs are potential distal regulatory elements (4.9-6.2% within 2.5 kb of a TSS). For DMRs, the vast majority (93.8%) are also located more than 2.5 kb away from a TSS. Across DMRs that show hypo-methylation in at least one INTACT sample (Figure 5D), between 36,643 and 83,992 are hypo-methylated in a single neuron subtype. Recapitulating the division of ATAC-seq peaks (Figure 5B), excitatory neurons have the highest number of hypo-DMRs (Figure 5D), and remarkably, most are not shared with PV or VIP neurons. Taken together, these data extend previous profiling experiments in the brain, first by identifying hundreds of thousands of putative regulatory regions across three neuron subtypes, and then by showing that distinct sets are active in individual subtypes.

E. Cell Type-Specific Hypo-Methylation at Activity-Induced Transcription Factor Binding Sites

Because regions bound by activity-dependent TFs, as a whole, show constitutive DNA hypo-methylation (Guo et al., 2011), DMRs could point to regulatory regions with cell type-specific responses to induced neuronal activity. Therefore, we addressed the relationship between DMRs and activity-dependent TF binding in excitatory neurons, reasoning that our overall findings would also be applicable to the two inhibitory subpopulations that are not easily obtainable in quantities required for TF ChIP-seq. We examined activity-dependent TF binding profiles using previously published ChIP-seq data from cortical cultures largely composed of immature excitatory neurons (Kim et al., 2010; Malik et al., 2014). For all tested TFs, the majority of activity-dependent TF binding sites (58.2-83.9%) overlap with excitatory neuron UMRs+LMRs (Figure 6B, left). However, only 1.4% of CREB and 10.8% of SRF binding sites overlap with excitatory neuron-specific hypo-DMRs, compared to 33.4-40.3% of AP-1 (FOS, FOSB, JUNB) and NPAS4 binding sites ($p<2\times10^{-38}$, Fisher's Exact Test, FET). In particular, activity-dependent binding sites for AP-1 factors and NPAS4 in cortical cultures are enriched in excitatory hypo-DMRs and depleted in PV-, VIP-, and glia-specific hypo-DMRs (Figures 6B-C). This analysis suggests that excitatory neuron-specific hypo-DMRs overlapping AP-1 and NPAS4 binding sites are a set of candidate regions that coordinate activity-dependent responses unique to excitatory neurons. Likewise, hypo-DMRs in PV and VIP neurons provide a resource for identifying AP-1 and NPAS4 targets that orchestrate distinct activity-dependent responses in inhibitory neurons (Spiegel et al., 2014).

F. Neuronal Subtypes Show Coordinated Epigenomic Differences

Epigenomic marks carry information about cell function, via their correlation with gene expression and gene regulatory regions, as well as cell development (Bird, 2002; Hon et al., 2013; Stadler et al., 2011; Thurman et al., 2012). Therefore, we first assessed whether the profiled epigenomic marks were well-correlated with each other and then applied epigenomic marks to quantify relationships across cell types and developmental stages.

Cell type-specific hypo-methylation in the CG context is coordinated with hypo-methylation in the CH context (Figures 5E, left two panels, and 6D) and increased chromatin accessibility (Figure 5E, third panel). Excitatory neuron hypo-DMRs are also enriched for histone modifications associated with active enhancers (H3K4me1 and H3K27ac) but not promoters (H3K4me3) (Figure 5E, right panel). Similarly, ATAC-seq levels in excitatory neurons are correlated with both H3K4me1 and H3K27ac at enhancers (Figure 6E) but demarcate TF binding sites with greater spatial resolution (Figure 5E, third panel versus right panel). Overlapping features derived from multiple assays (Figure 6F) provide convergent evidence for identifying candidate regulatory regions, and both raw and processed data can be explored via a web-based browser (http://neomorph.salk.edu/mm_intact/).

We quantified the epigenomic relationships across cell types in several ways: by the similarity of DNA methylation patterns in 500 bp bins genome-wide (Figure 8A) and at ATAC-seq peaks (Figures 7A and 8B), and by the similarity of Tn5 insertion densities (Figure 7B) at ATAC-seq peaks. As expected, excitatory and NeuN+ neurons are strongly correlated using DNA methylation signal at both genomic bins and ATAC-seq peaks (Pearson r ~0.9), and hierarchical clustering groups excitatory neurons with NeuN+ neurons. PV and VIP neurons cluster together, in line with their functional roles as inhibitory neurons. In contrast, excitatory and VIP neurons show the lowest similarity across INTACT-purified cell types. Unexpectedly, DNA methylation levels in fetal brain and in glia correlate more strongly with VIP neurons than with excitatory or PV neurons. At ATAC-seq peaks (Figures 7A and 8B), this similarity among VIP, fetal, and glial samples could suggest that more gene regulatory characteristics of immature neurons are retained by VIP neurons than by excitatory or PV neurons. Collectively, our data demonstrates that DNA methylation and chromatin features reveal a coordinated, hierarchical organization of mature cortical cell types that is reflected across much of the genome.

G. Distinct Sets of DNA Binding Factors Act at Putative Neuron Subtype-Specific Regulatory Regions

We next sought to characterize the DNA binding TFs that are responsible for these unique neuronal regulatory landscapes. Our RNA-seq analysis identified 267 differentially expressed TFs. These include TFs that play well-known regulatory roles in the development of each cell type (e.g., *Lhx6* in PV interneurons and *Prox1* in VIP interneurons) (Kessaris et al., 2014) as well as many other TFs with unknown neuronal functions.

TF binding enhances chromatin accessibility, but the central region of binding is protected from the activity of enzymes such as Tn5 transposase, resulting in a notch, or footprint, in the ATAC-seq profile (Buenrostro et al., 2013). In agreement with previous footprinting studies (Neph et al., 2012), we observe a range of footprint shapes for different TFs (Figure 9A). With the notable exceptions of CTCF and ZFP410, footprinted sites in a cell are generally associated with reduced regional DNA methylation levels (Figure 10A).

We applied footprint analysis of ATAC-seq datasets to infer TF binding at cell type-specific regulatory regions and combined it with complementary analysis of DNA binding sequence motifs enriched at hypo-DMRs. We focused on footprints and motifs of moderately to highly expressed TFs (TPM≥30) and identified 68 TFs that may regulate cell type-specific gene expression (Figure 9B). Overall, both our footprint and motif predictions converge on similar sets of enriched and depleted TFs. These sets encompass both well-established and novel TFs. In excitatory neurons, both footprint and motif predictions show overrepresentation of *Egr*, AP-1 family members, *Neurod2*, *Rfx1/3/5*, and *Tbr1*. Two TF groups potentially linked to PV neuron development, *Mafb/g* and *Mef2a/c/d* (Kessaris et al., 2014), are among those enriched in PV-specific footprints and hypo-DMRs (Figure 9B) as well as PV hypo-DMRs shared with both excitatory and VIP neurons (Figure 10B). Studies of MEF2 have largely focused on its role in excitatory neurons (Rashid et al., 2014); here, both footprinting and motif analyses suggest a critical function for MEF2 in PV neurons at PV-specific regulatory regions. Interestingly, VIP neuron footprints and DMRs are enriched for TFs best known for their developmental roles (e.g., *Dlx*, *Pou*, and *Sox* family members; *Arx* and *Vax2*) (Kessaris et al., 2014), an extension of our previous observation that VIP methylomes share common patterns with fetal and glial methylomes. Motifs for these TFs are also enriched at fetal and glial hypo-DMRs, including those that are shared with VIP neurons (Figure 10B).

TFs control complex cellular processes by forming networks of mutual regulation, yet differences in TF regulatory networks between neuron types are largely unknown. We examined regulatory interactions among TFs by building networks of predicted cell type-specific TF regulation, as well as a pan-neuronal regulatory network (Figures 9C-D). These networks recover a number of previously implicated TF-TF regulatory interactions and suggest novel interactions. For example, our prediction that MEF2D targets *Dlx6* in PV neurons parallels the requirement of a homolog, MEF2C, for *Dlx6* expression in branchial arches (Verzi et al., 2007).

To explore the potential contribution of ATAC-seq peaks and footprints to the regulation of nearby gene expression, we examined their coverages around the TSS of highly cell type-specific genes. Differentially expressed genes display an increased density of cell type-specific footprints centered around the TSS (Figure 9E) and are significantly enriched for cell type-specific ATAC-seq peaks (Figure 9F). When we examined pan-neuronal genes (Hobert et al., 2010) such as *Pclo*, *Rims1*, *Cdh2*, and *Grip1* (Figure 10C), we noted that they were also surrounded by an array of ATAC-seq peaks, many of which were active exclusively in one neuron class. Indeed, we find that cell type-specific ATAC-seq peaks are moderately enriched around the TSS of pan-neuronal genes (Figure 9F), highlighting the potential for these regions to shape neuronal identity by regulating both cell type-specific and pan-neuronal programs of gene expression.

H. Among DNA Methylation and Chromatin Accessibility Features, Non-CG Methylation Best Correlates With RNA Abundance

Genome-wide, we find a strong correlation between RNA abundance and both DNA methylation and ATAC-seq signals around the TSS (Figure 11A). For both mCG and mCH, the correlation extends throughout the gene body, with a peak ~1-2 kb downstream of the TSS. At differentially expressed genes, mCH is significantly more correlated with expression (Spearman r=0.50) than mCG (r=0.34; p=0.0063, t-test using the three cell types as samples) or ATAC-seq insertion density (r=0.25; $p=5.4\times10^{-4}$). A generalized linear model with a sparseness-promoting regularization (LASSO) using mCG, mCH, and ATAC-seq features further identifies gene body mCH as the most informative feature for inferring RNA abundance (Figures 12A-B).

Our finding that the strongest correlation between RNA levels and mCG occurs ~1-2 kb downstream of the TSS agrees with recent findings in medulloblastoma cell lines (Hovestadt et al., 2014) and in human cardiomyocytes (Gilsbach et al., 2014). Our results extend this observation to mCH and show that mCH, an epigenetic modification abundant across diverse classes of neocortical neurons, is better correlated with gene abundances measured by RNA-seq. Future studies using more direct measures of gene transcription are warranted to complement these findings.

I. Gene Clusters Based on Intragenic Non-CG Methylation Share Gene Expression, Chromatin, and Functional Organization

As described above, non-CG methylation within the gene body is inversely correlated with gene expression. Yet, this epigenomic feature may display greater divergence across neuron types than their transcriptional configurations (Figures 3B, F), suggesting that it contains additional information related to cell type-specific differences. To explore this idea, we used an unbiased clustering approach to group genes by their patterns of intragenic mCH, followed by an integrative analysis of gene expression, chromatin accessibility, and gene ontology. 23,023 genes were grouped into 25 clusters by their levels of intragenic mCH, normalized by the level in the flanking region (Figures 11B-E and 12C). Approximately half of these genes share similar patterns of mCH across neurons, including hyper-methylated genes with low expression levels (clusters 2,6; 13.5% of genes) and hypo-methylated genes with moderate to high expression (clusters 3-5,7-8; 40%). The latter category is not enriched for differentially expressed genes (Figure 11D) but is enriched for cell type-specific ATAC-seq peaks (Figure 11E). By gene ontology (GO) analysis (Huang et al., 2009), genes in these clusters tend to be enriched for general cellular processes, for example, transcription (GO:0006350) and RNA binding (GO:0003723).

The remaining half of genes capture the spectrum of intragenic CH methylation across neuronal populations by clustering into groups showing neuron subtype-specific hyper- and hypo-mCH. Clusters 10-18 (23.6% of genes) are hyper-methylated at CH sites in one or more cell types and are expressed at relatively low levels. Clusters 19-25 (17.8% of genes) are hypo-methylated in specific cell types and are generally expressed at higher levels, with hypo-methylation occurring together with increased expression (e.g., Cluster 22 enriched for PV>Exc and VIP genes). These clusters are enriched for both differentially expressed genes and accessible chromatin. Although genes that are exclusively expressed in only one or two cell types are grouped in clusters 19-25, a subset of pan-neuronal genes that differ in their expression levels across neuronal subtypes are also grouped here (e.g., *Cdh2*, *Grip1*, *Bsn*). These clusters also contain pan-neuronal genes that do not meet our threshold for differential expression (e.g., *Anks1b*), an example of the ability of intragenic mCH to parse the neuronal transcriptome into finer patterns.

Several clusters with cell type-specific hypo-methylation are enriched for neuronal GO categories, for example, postsynaptic density (GO:0014069: 6.7-fold enrichment, q=0.035, cluster 19) and synapse (GO:0045202: 2.6-fold, q=0.033, cluster 20; 2.9-fold, $q=2.8\times10^{-4}$, cluster 21). Neuron subtype-specific differences in intragenic mCH may be especially relevant in light of recent evidence that MeCP2 binding to mCA represses transcription of long neuronal genes (Gabel et al., 2015). The enrichment of neuronal GO categories at these clusters suggests that cell type-specific expression levels of genes with neuronal functions may partly be a consequence of differences in levels of intragenic mCH.

J. Non-CG Methylation is Lowest at the Nucleosome Center and Increases at Linker Regions

In addition to its variations with gene expression, we asked if mCH levels also differed relative to chromatin features such as nucleosome positioning. We estimated nucleosome locations using ATAC-seq and found that coherently phased modulation of mCH is evident over approximately 1 kb (~5 nucleosomes), decreasing by up to 9.5% at the nucleosome center and increasing by 11.1% in neighboring linkers (Figure 11F). mCG levels display a similar but weaker modulation (<2%) (Figure 12D). Our results support earlier studies in the CG context (Teif et al., 2014) and extend the link between nucleosome positioning and DNA methylation in mammalian cells to the non-CG context.

K. **Identification of Distinct Classes of Large Hypo-Methylated Regions**

We further sought to identify multi-kilobase regions of low DNA methylation in our datasets. Hypo-DMRs are not randomly distributed in the genome, but instead show a bimodal distribution of inter-DMR distances (Figure 13A). Closely-spaced hypo-DMRs may represent fragments of larger hypo-methylated features. Therefore, we merged neuron subtype-specific hypo-DMRs located within 1 kb of each other and defined those exceeding 2 kb in length as "large hypo-DMRs" (Figure 13B, left). We also observed another category of large hypo-methylated domains that are consistent with previously described DNA methylation valleys (DMVs) or canyons (Jeong et al., 2014; Xie et al., 2013) (Figure 13B).

Although both are multi-kilobase hypo-methylated regions, large hypo-DMRs and DMVs occupy distinct genomic locations (Figure 14A). Compared to large hypo-DMRs, DMVs have higher overlap across cell types (Figure 14A) and more extreme lengths (Figure 14B), extending up to 104 kb compared to large hypo-DMRs, which extend to 32 kb. Consistent with their higher GC content (Figure 14C) and lower levels of CG methylation (Figure 14D), most DMVs (85-94%) overlap CpG islands. In contrast, only 1-9% of large hypo-DMRs overlap CpG islands. Furthermore, DMVs straddle the TSS whereas large hypo-DMRs are enriched downstream of the TSS (Figure 14E).

To better characterize different classes of hypo-methylated regions, we took advantage of our histone modification data in excitatory neurons. Large hypo-DMRs show higher levels of histone modifications associated with active enhancers, H3K27Ac and H3K4me1, compared to DMRs <2 kb (Figure 13C, left). Excitatory DMVs display a bimodal distribution for H3K4me3 and H3K27me3 and can be divided as H3K4me3+ (Figure 13B, left) versus H3K27me3+ (Figure 13B, right). As expected, H3K27me3+ DMVs are depleted for ATAC-seq reads and overlap genes with low expression (Figure 13C, middle and right). Large hypo-DMRs and H3K4me3+, but not H3K27me3+, DMVs are enriched for differentially expressed genes (Figure 13D). In fact, the bimodal distribution of H3K4me3 and H3K27me3 levels in DMVs suggests that these domains can be associated with either active or repressed genes, and the two histone modifications partition DMVs into functionally distinct categories (Figures 14F-G).

L. Hyper-Methylation at Cell Type-Specific Transcription Factor Genes Preserves a Trace of Early Developmental Expression

DMVs are highly overlapping across adult cell types and fetal cortex (Figure 14A), in line with previous evidence (Xie et al., 2013) suggesting they may be established early during development and subsequently maintained. To address whether these regions are dynamically modified during development, we compared the boundaries of fetal DMVs between fetal and adult cells. Genome-wide, 51-67% of fetal DMVs remain as DMVs in adult neurons and glia but gain methylcytosines, resulting in a contraction of DMV length as the brain matures (median decrease = 747 bp; $p<2\times10^{-18}$, Wilcoxon rank sum).

We further focused our analysis on fetal DMVs overlapping genes. Fetal DMVs are highly enriched for TF genes (Figure 14G), and 75 out of 77 fetal DMVs associated with a list of candidate developmental TFs (Visel et al., 2013) are shorter in at least one adult cell type (Figure 14H). To identify the DMVs that display the most significant developmental mCG gains, we compared mCG levels across fetal and adult cells in the interior of fetal DMVs; to avoid the possible confound of intragenic DNA methylation, we used the DMV interior upstream of the TSS (Figure 13E). This analysis identified 454 genes (66%; FET, $q<0.01$) that exhibit significantly increased mCG in at least one adult cell type versus fetal cortex; 210 genes (31%) have more than a 5-fold increase.

When we examined the genes with the highest increases in mCG, we noted that several code for critical TFs known to shape neuronal subtype identity and are predominantly expressed in neural progenitor cells and immature precursors. At these TF loci, the highest mCG fold change generally occurs in the cell type where the gene is active in development but down-regulated in the adult. For example, *Neurog2* is highly expressed during embryonic development in the common progenitors of cortical excitatory neurons and many glial cells, but it is not expressed in these cells in the adult brain nor at any time during inhibitory neuron development (Sommer et al., 1996; Wang et al., 2013). Our DNA methylation data shows that *Neurog2* lies within a DMV in all cells except excitatory neurons and glia, where the region is hyper-methylated (Figure 13F, left). In contrast, *Nkx2-1* is specifically expressed in the medial ganglionic eminence (MGE), the birthplace of cortical PV neurons (DeFelipe et al., 2013). Immature cortical PV neurons switch off *Nkx2-1* soon after leaving the MGE in order to direct their migration to the cortex; neurons that maintain *Nkx2-1* expression instead travel to the striatum (Nóbrega-Pereira et al., 2008). An extended (>15 kb) DMV covers *Nkx2-1* in fetal cortex, excitatory neurons, VIP neurons, and glia, yet this DMV is only ~6.5 kb in PV neurons (Figure 13F, right). Similar findings are seen at DMVs overlapping *Dlx2*, *Pax6*, *Vax1*, and *Gsx2* (Figures 14I-J).

At these TF loci, the methylomes of adult neurons contain a signature of past gene expression. In contrast to the rest of the genome, hyper-methylation, rather than hypo-methylation, marks the relevant cell type-specific genes. In contrast to vestigial enhancers (Hon et al., 2013), this epigenetic trace of the neuron's development arises from the gain of cell type-specific hyper-methylation rather than the retention of hypo-methylation. We further asked what fraction of this hyper-methylation is a result of hmC rather than mC. For DMVs at *Neurog2* and *Pax6*, we find that adult frontal cortex hmCG levels from TAB-seq (Lister et al., 2013) are approximately 10% of excitatory neuron MethylC-seq signals at CG sites. Because we lack matched hmC data from purified excitatory neurons, the precise contribution of hmCG is difficult to assess, although we believe from this comparison that the majority of the hyper-methylation originates from mCG. Furthermore, at non-CG sites in these two DMVs, we find that essentially all of the observed hyper-methylation originates from mCH, consistent with evidence that hydroxymethylation occurs nearly exclusively in the CG context (Yu et al., 2012).

Chapter IV

Discussion

This study introduces the INTACT system in mice, the first method to affinity purify nuclei from genetically-defined cell types in a mammal. INTACT efficiently isolates nuclei from both common and rare cell types, enabling us to examine the epigenomic organization of neocortical excitatory, PV, and VIP neurons with unprecedented cell type-specific resolution. We find that the morphological and physiological diversity of neocortical neurons is paralleled by widespread differences in their underlying epigenomes. By using coordinated epigenomic marks to show that neocortical neurons adopt unique regulatory landscapes, our data adds a new resource to existing catalogues of transcriptional diversity. We further identify candidate TFs acting at regulatory regions and demonstrate how epigenomic states of adult cells capture long-lasting attributes of neuronal identity, including patterns of past gene expression, current gene expression, and potential experience-dependent responses. In particular, we find a close relationship between intragenic non-CG methylation and differential gene expression. Furthermore, purified neuronal epigenomes reveal distinctive hyper-methylation patterns associated with developmentally transient expression of critical TFs that shape neuronal subtype identity.

Affinity Purification of Nuclei Facilitates Epigenomic Studies

INTACT is uniquely suited to investigating cell type-specific epigenomes, an application that can be challenging with other purification methods. Genome-wide epigenomic assays generally require tens of thousands to millions of cells, which limit the utility of manual sorting for this purpose. Methods that involve cellular dissociation in the adult brain may be inefficient and induce stress responses that alter the cellular state. In contrast, INTACT couples rapid tissue lysis with gentle isolation of sufficient numbers of cell type-specific nuclei for epigenomic studies. Whereas FACS-sorted cells or nuclei may be fragile and difficult to manipulate, the attachment of magnetic beads to nuclei in INTACT greatly simplifies buffer exchanges and volume reductions. Furthermore, unlike FACS or laser capture microdissection, INTACT requires no specialized instruments. INTACT is particularly well-suited for isolating rare cell types; cells constituting 1-3% of the starting material can be enriched to >98% purity and subsequently used for MethylC-seq and ATAC-seq.

In this study, we have focused on cellular diversity in the healthy mammalian brain. INTACT can also be used to explore cell type-specific epigenomics in mouse models of schizophrenia, autism, neurodegeneration, and other neuropsychiatric disorders, or adapted for use in non-neuronal tissues. In addition to epigenomic studies, INTACT is an efficient method for isolating nuclear RNA from defined cell types that complements existing strategies for RNA profiling. We note that some degree of non-specific RNA contamination is intrinsic to affinity purification strategies, including INTACT. Nevertheless, we have shown that INTACT expression profiles recover known cell markers and can be used to discover novel markers.

Cell Type-Specific Developmental Signatures are Encoded in the Methylomes of Adult Cells

Mature neuronal diversity arises from a developmental odyssey. Whereas one class of large hypo-methylated regions (large hypo-DMRs) reflects the neuron’s current transcriptional state, a second class (DNA methylation valleys, DMVs) reveals patterns of past gene expression. We find that a subset of genes coding for TFs that establish neuronal identity, including *Neurog2*, *Nkx2-1*, *Dlx2*, *Pax6*, *Vax1*, and *Gsx2*, overlap with DMVs showing cell type-specific hyper-methylation. At these genes, hyper-methylation at DMVs in the adult methylome provides a record of transient high TF expression during development, whereas the same genes are hypo-methylated in other cell types. We speculate that this pattern might arise if (1) these DMVs are initially marked by H3K27me3 in neural progenitors (Xie et al., 2013), (2) H3K27me3 is removed in a particular neuronal lineage to allow TF expression at the appropriate developmental time point, and (3) this removal simultaneously increases the accessibility of the region to DNA methyltransferases, whereas other cell types maintain an inaccessible chromatin state throughout development and into adulthood. Measuring gene expression in defined populations of immature cells can be challenging as they are intermixed and often do not express the terminal markers of adult neuronal subtypes. Our data suggests that developmental TF expression could be predicted from DNA methylation patterns in adult cells, providing an alternate approach to investigating cell type-specific developmental history.

Genome-Wide Analyses Parse Neuronal Diversity

Neuronal cell types have been defined based on morphology, electrophysiology, connectivity, and, more recently, patterns of gene expression and regulation. Traditional approaches for investigating these features produce datasets of modest size and with a relatively small number of independent parameters, which limit the distinctions that can be made among neuronal cell types. As demonstrated here, genome-wide approaches generate large and information-rich datasets that reveal complex neuron subtype-specific patterns of transcript abundances, DNA methylation, and chromatin accessibility. Genome-wide information derived from these datasets can be used to parse neuronal subtypes into even finer divisions based on patterns of both gene expression and gene regulation, which in turn can be combined with transgenic approaches to label new subpopulations of neurons and enable their purification. The synergy between genetic engineering of experimental organisms, cell type-specific purification, and genome-scale data analysis promises a new and comprehensive view of neuronal diversity in the mammalian brain.

Chapters I-IV of this dissertation has been submitted to Neuron for consideration.

Author contributions

J.N., J.R.E., S.R.E., and T.J.S. designed and supervised the research. A.M. designed and generated the INTACT mouse, developed the affinity purification, and performed nuclei isolations and related experiments. G.L.H. designed the Sun1 tag and constructed RNA-seq libraries. G.L.H. and S.P. constructed ATAC-seq libraries. A.M. and G.L.H. constructed ChIP-seq libraries. S.P. sequenced the RNA-seq, ChIP-seq, and ATAC-seq libraries. M.A.U constructed MethylC-seq libraries. J.R.N. sequenced and mapped the MethylC-seq libraries. R.L. provided initial analysis for the MethylC-seq data. E.A.M., F.P.D., C.L., A.M., and R.L. analyzed data. A.M., E.A.M., and J.N. prepared the manuscript. F.P.D., C.L., J.R.E., S.R.E., G.L.H., and R.L. revised the manuscript. E.A.M., F.P.D., C.L., and A.M. are equally responsible for the analysis results. A.M., E.A.M., F.P.D., and C.L. contributed equally to this work.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute (J.R.E., S.R.E., J.N., and

T.J.S.), 1-U01-MH105985 (J.R.E.), MSTP training grant (A.M.), and NIH/NINDS R00NS080911 (E.A.M.). J.N., S.R.E., and T.J.S. are Howard Hughes Medical Institute Investigators. J.R.E. is a Howard Hughes Medical Institute and Gordon and Betty Moore Investigator. We thank Hao Zhang at the Johns Hopkins School of Public Health for flow cytometry. We thank Yupeng He for comments and suggestions.

References

Amin, N.M., Greco, T.M., Kuchenbrod, L.M., Rigney, M.M., Chung, M.I., Wallingford, J.B., Cristea, I.M., and Conlon, F.L. (2014). Proteomic profiling of cardiac tissue by isolation of nuclei tagged in specific cell types (INTACT). Development *141*, 962-973.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202-208.

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

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213-1218.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. *23*, 341-351.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev. Cell *18*, 1030-1040.

DeFelipe, J., López-Cruz, P.L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., Cauli, B., Fairén, A., Feldmeyer, D., et al. (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat. Rev. Neurosci. *14*, 202-216.

Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F., Stevens, T.R., Ma, G., Bupp, S., Shrestha, P., Shah, R.D., Doughty, M.L., et al. (2008). Application of a translational profiling approach for the comparative analysis of CNS cell types. Cell *135*, 749-762.

Dugas, J.C., Tai, Y.C., Speed, T.P., Ngai, J., and Barres, B.A. (2006). Functional genomic analysis of oligodendrocyte differentiation. J. Neurosci. *26*, 10967-10983.

Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A., and Liotta, L.A. (1996). Laser capture microdissection. Science *274*, 998-1001.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1-22.

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature doi: 10.1038/nature14319.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810-822.

Gay, L., Miller, M.R., Ventura, P.B., Devasthali, V., Vue, Z., Thompson, H.L., Temple, S., Zong, H., Cleary, M.D., Stankunas, K., et al. (2013). Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. Genes Dev. *27*, 98-115.

Gelman, D.M., and Marín, O. (2010). Generation of interneuron diversity in the mouse cerebral cortex. Eur. J. Neurosci. *31*, 2136-2141.

Gilsbach, R., Preissl, S., Grüning, B.A., Schnick, T., Burger, L., Benes, V., Würch, A., Bönisch, U., Günther, S., Backofen, R., et al. (2014). Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease. Nat. Commun. *5*, 5288.

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

Guo, J.U., Ma, D.K., Mo, H., Ball, M.P., Jang, M.H., Bonaguidi, M.A., Balazer, J.A., Eaves, H.L., Xie, B., Ford, E., et al. (2011). Neuronal activity modifies the DNA methylation landscape in the adult brain. Nat. Neurosci. *14*, 1345-1351.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat. Neurosci. *17*, 215-222.

Heiman, M., Schaefer, A., Gong, S., Peterson, J.D., Day, M., Ramsey, K.E., Suárez-Fariñas, M., Schwarz, C., Stephan, D.A., Surmeier, D.J., et al. (2008). A translational profiling approach for the molecular characterization of CNS cell types. Cell *135*, 738-748.

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

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311-318.

Henry, G.L., Davis, F.P., Picard, S., and Eddy, S.R. (2012). Cell type-specific genomics of Drosophila neurons. Nucleic Acids Res. *40*, 9691-9704.

Hobert, O., Carrera, I., and Stefanakis, N. (2010). The molecular and gene regulatory signature of a neuron. Trends Neurosci. *33*, 435-445.

Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat. Genet. *45*, 1198-1206.

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

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44-57.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G.A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., et al. (2014). Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat. Genet. *46*, 17-23.

Jiang, Y., Matevossian, A., Huang, H.S., Straubhaar, J., and Akbarian, S. (2008). Isolation of neuronal chromatin from brain tissue. BMC Neurosci. *9*, 42.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. *42*, D764-770.

Kepecs, A., and Fishell, G. (2014). Interneuron cell types are fit to function. Nature *505*, 318-326.

Kessaris, N., Magno, L., Rubin, A.N., and Oliveira, M.G. (2014). Genetic programs controlling cortical interneuron fate. Curr. Opin. Neurobiol. *26*, 79-87.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168-176.

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics *29*, 1035-1043.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Lister, R., O’Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

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

Ma, H., Morey, R., O'Neil, R.C., He, Y., Daughtry, B., Schultz, M.D., Hariharan, M., Nery, J.R., Castanon, R., Sabatini, K., et al. (2014). Abnormalities in human pluripotent cells due to reprogramming mechanisms. Nature *511*, 177-183.

Malik, A.N., Vierbuchen, T., Hemberg, M., Rubin, A.A., Ling, E., Couch, C.H., Stroud, H., Spiegel, I., Farh, K.K., Harmin, D.A., et al. (2014). Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. Nat. Neurosci. *17*, 1330-1339.

Maze, I., Shen, L., Zhang, B., Garcia, B.A., Shao, N., Mitchell, A., Sun, H., Akbarian, S., Allis, C.D., and Nestler, E.J. (2014). Analytical tools and current challenges in the modern era of neuroepigenomics. Nat. Neurosci. *17*, 1476-1490.

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

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

Micallef, L., and Rodgers, P. (2014). eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. PLoS One *9*, e101717.

Molyneaux, B.J., Arlotta, P., Menezes, J.R., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat. Rev. Neurosci. *8*, 427-437.

Molyneaux, B.J., Goff, L.A., Brettler, A.C., Chen, H., Brown, J.R., Hrvatin, S., Rinn, J.L., and Arlotta, P. (2014). DeCoN: Genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex. Neuron *85*, 275-288.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Nóbrega-Pereira, S., Kessaris, N., Du, T., Kimura, S., Anderson, S.A., and Marín, O. (2008). Postmitotic Nkx2-1 controls the migration of telencephalic interneurons by direct repression of guidance receptors. Neuron *59*, 733-745.

Okaty, B.W., Miller, M.N., Sugino, K., Hempel, C.M., and Nelson, S.B. (2009). Transcriptional and electrophysiological maturation of neocortical fast-spiking GABAergic interneurons. J. Neurosci. *29*, 7040-7052.

Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nat. Biotechnol. *24*, 79-88.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447-455.

Pohl, A., and Beato, M. (2014). bwtool: a tool for bigWig files. Bioinformatics *30*, 1618-1619.

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

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187-191.

Rashid, A.J., Cole, C.J., and Josselyn, S.A. (2014). Emerging roles for MEF2 transcription factors in memory. Genes Brain Behav. *13*, 118-125.

Rudy, B., Fishell, G., Lee, S., and Hjerling-Leffler, J. (2011). Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. Dev. Neurobiol. *71*, 45-61.

Sanz, E., Yang, L., Su, T., Morris, D.R., McKnight, G.S., and Amieux, P.S. (2009). Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. Proc. Natl. Acad. Sci. USA *106*, 13939-13944.

Saxena, A., Wagatsuma, A., Noro, Y., Kuji, T., Asaka-Oba, A., Watahiki, A., Gurnot, C., Fagiolini, M., Hensch, T.K., and Carninci, P. (2012). Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. Biotechniques *52*, 381-385.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., et al. (2013). Patterns of population epigenomic diversity. Nature *495*, 193-198.

Sommer, L., Ma, Q., and Anderson, D.J. (1996). Neurogenins, a novel family of atonal-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor cell heterogeneity in the developing CNS and PNS. Mol. Cell. Neurosci. *8*, 221-241.

Soriano, P. (1999). Generalized lacZ expression within the ROSA26 Cre reporter strain. Nat. Genet. *21*, 70-71.

Spiegel, I., Mardinly, A.R., Gabel, H.W., Bazinet, J.E., Couch, C.H., Tzeng, C.P., Harmin, D.A., and Greenberg, M.E. (2014). Npas4 regulates excitatory-inhibitory balance within neural circuits through cell-type-specific gene programs. Cell *157*, 1216-1229.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.

Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al.; Mouse ENCODE Consortium. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Steiner, F.A., Talbert, P.B., Kasinathan, S., Deal, R.B., and Henikoff, S. (2012). Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. Genome Res. *22*, 766-777.

Sugino, K., Hempel, C.M., Miller, M.N., Hattox, A.M., Shapiro, P., Wu, C., Huang, Z.J., and Nelson, S.B. (2006). Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat. Neurosci. *9*, 99-107.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. *13*, 537-551.

Teif, V.B., Beshnova, D.A., Vainshtein, Y., Marth, C., Mallm, J.P., Höfer, T., and Rippe, K. (2014). Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. Genome Res. *24*, 1285-1295.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Verzi, M.P., Agarwal, P., Brown, C., McCulley, D.J., Schwarz, J.J., and Black, B.L. (2007). The transcription factor MEF2C is required for craniofacial development. Dev. Cell *12*, 645-652.

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R.V., McKinsey, G.L., Pattabiraman, K., Silberberg, S.N., Blow, M.J., et al. (2013). A high-resolution enhancer atlas of the developing telencephalon. Cell *152*, 895-908.

Wang, B., Long, J.E., Flandin, P., Pla, R., Waclaw, R.R., Campbell, K., and Rubenstein, J.L. (2013). Loss of Gsx1 and Gsx2 function rescues distinct phenotypes in Dlx1/2 mutants. J. Comp. Neurol. *521*, 1561-1584.

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

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell *153*, 1134-1148.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149*, 1368-1380.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zhang, H.M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.Y. (2012). AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. *40*, D144-149.

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

Figure 1. An Affinity Purification Method Isolates Cell Type-Specific Nuclei in Mice

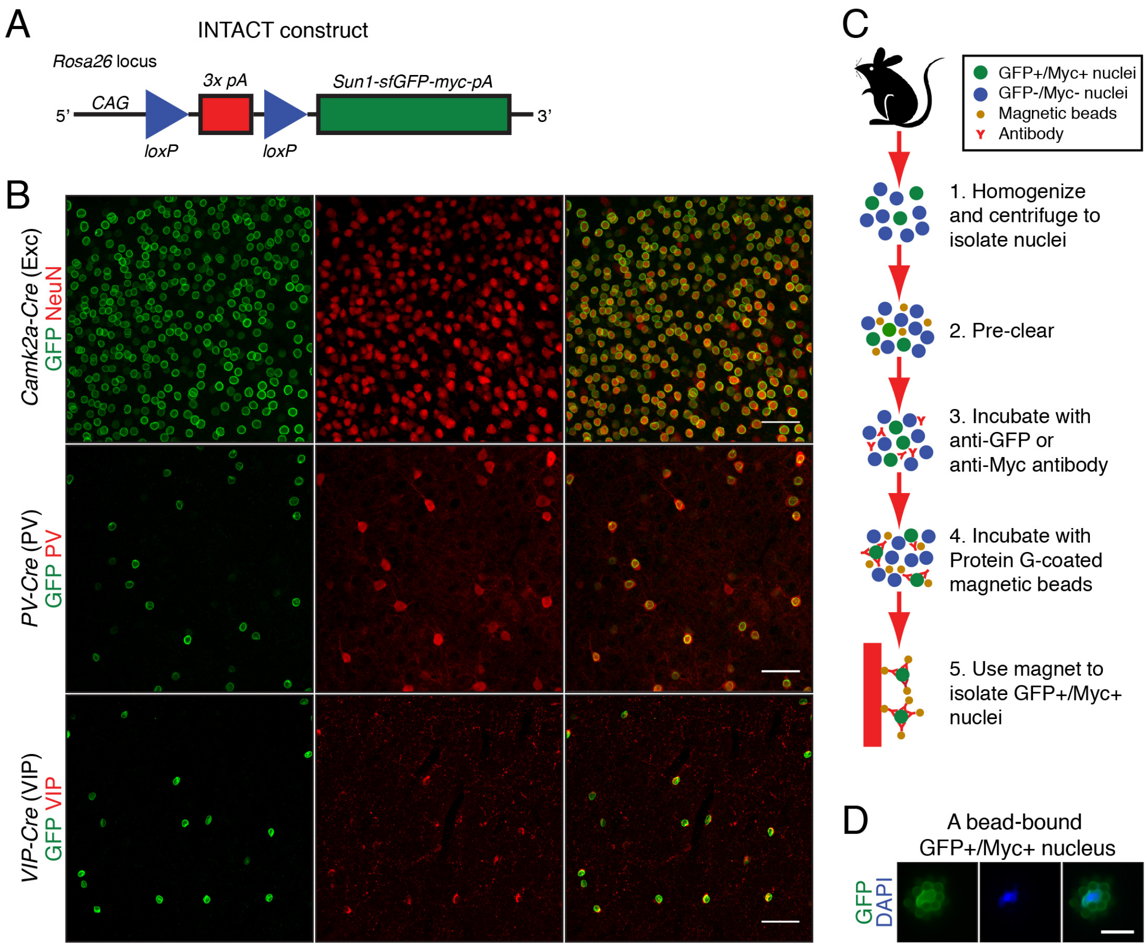
(A) Diagram of the INTACT knock-in mouse construct. Cre-mediated excision of the transcription stop signals activates expression of the nuclear membrane tag (*Sun1-sfGFP-myc*) in the cell type of interest.

(B) Immunohistochemistry showing localization of SUN1-sfGFP-Myc in neocortical excitatory, PV, and VIP neurons in mice that carry *R26-CAG-LSL-Sun1-sfGFP-myc* together with a Cre driver. Scale bars: 50 µm.

(C) Steps in the affinity purification method (INTACT).

(D) An example of a GFP+/Myc+ nucleus bound by Protein G-coated magnetic beads following INTACT purification and staining with DAPI. Scale bar: 10 µm.

(E) For each experiment, INTACT purifications were performed with anti-GFP using pooled neocortices of two mice. Specificity of mouse INTACT: after INTACT purification, bead-bound nuclei were stained with DAPI, and the numbers of GFP+ versus GFP- nuclei were quantified by fluorescence microscopy (100-200 nuclei per experiment). Yield of mouse INTACT: the total number of input nuclei, the % of GFP+ nuclei in the input, and the total number of bead-bound nuclei after INTACT purification were quantified using fluorescence microscopy or a hemocytometer (100-200 nuclei per experiment). The yield was calculated based on the observed number of bead-bound nuclei versus the expected number from the input. For % GFP+ nuclei in the input, the mean is shown. For quantities after INTACT purification, both the mean and ranges are shown.


A
INTACT construct
Rosa26 locus
5'
CAG
loxP
3x pA
loxP
Sun1-sfGFP-myc-pA
3'
B
Camk2a-Cre (Exc)
GFP NeuN
PV-Cre (PV)
GFP PV
VIP-Cre (VIP)
GFP VIP
C
GFP+/Myc+ nuclei
GFP-/Myc- nuclei
Magnetic beads
Antibody
1. Homogenize and centrifuge to isolate nuclei
2. Pre-clear
3. Incubate with anti-GFP or anti-Myc antibody
4. Incubate with Protein G-coated magnetic beads
5. Use magnet to isolate GFP+/Myc+ nuclei
D
A bead-bound GFP+/Myc+ nucleus
GFP
DAPI

E

INTACT specificity and yield

	Input nuclei	Nuclei after INTACT purification (anti-GFP)		
Cre driver	**% GFP+**	**Specificity (% GFP+)**	**# of isolated nuclei**	**Yield**
Camk2a-Cre (n=5)	50.9%	98.5% (95.0 - 100%)	3.13 mill (1.2 - 3.7 mill)	67.6% (48.2 - 88.4%)
PV-Cre (n=4)	2.8%	98.2% (95.6 - 100%)	146,042 (83,250 - 220,150)	59.6% (38.9 - 80.0%)
VIP-Cre (n=5)	1.6%	98.6% (95.7 - 100%)	92,042 (69,000 - 100,267)	56.2% (36.9 - 74.0%)

Figure 2. Nuclear Labeling and Specificity of Mouse INTACT

(A) Hippocampus, kidney, and heart from adult *Sox2-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice, where Cre recombination occurs at the early embryo stage. Scale bar: 50 µm.

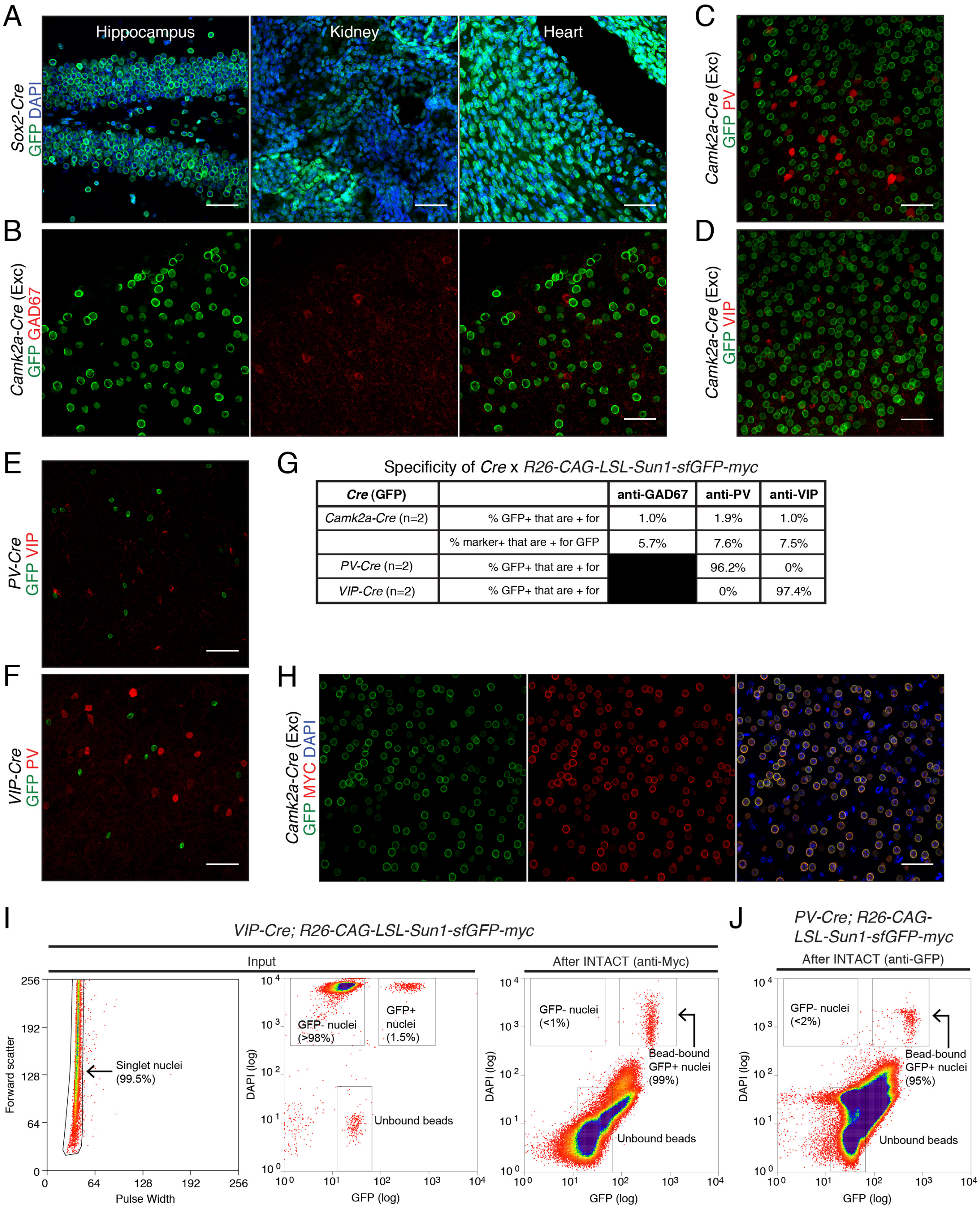
(B) Immunohistochemistry showing that neocortical GFP+ nuclei in *Camk2a-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* adult mice do not co-localize with GAD67, an inhibitory neuron marker. Scale bar: 50 µm.

(C-F) Immunohistochemistry showing that neocortical GFP+ nuclei in *Camk2a-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* adult mice do not co-localize with PV (C) or VIP (D). Similarly, neocortical GFP+ nuclei in *PV-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* adult mice do not co-localize with VIP (E), and neocortical GFP+ nuclei in *VIP-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* adult mice do not co-localize with PV (F). Scale bars: 50 µm.

(G) Quantification of Cre driver specificity by immunostaining. Each Cre driver was crossed with *R26-CAG-LSL-Sun1-sfGFP-myc* mice, and the percent of GFP+ cells that co-localize with the indicated markers was quantified. For Camk2a-Cre driver, the percentage of GAD67, PV, and VIP cells that co-localize with GFP was also quantified; furthermore, quantification of GFP and NeuN staining (Figure 1B) showed that 100% of GFP+ nuclei were also NeuN+. Counts were made using 100 µm vibratome sections and >200 nuclei per mouse (n=2).

(H) Myc labeling co-localizes with GFP labeling in the neocortex of adult *Camk2a-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice. Scale bar: 50 µm.

(I and J) Assessment of INTACT purification by flow cytometry. INTACT was performed using anti-Myc to isolate nuclei from the neocortices of two *VIP-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice (I). Analysis of input nuclei (after step 2 in Figure 1C) shows that 99.5% of input nuclei are singlets (left), and 1.5% of input nuclei are GFP+ (middle). Unbound beads remaining after the pre-clear step were identified using a beads-only control (data not shown). After INTACT purification, 99% of bead-bound nuclei are GFP+ (right). Because multiple magnetic beads are bound to each GFP+/Myc+ nucleus, the DAPI fluorescence is variably reduced relative to input nuclei. INTACT was performed using anti-GFP to isolate nuclei from the neocortices of two *PV-Cre; R26-CAG-LSL-Sun1-sfGFP-myc* mice (J). After INTACT purification, 95% of bead-bound nuclei are GFP+. The percentages of singlet nuclei, GFP- nuclei, and GFP+ nuclei were determined by the gates outlined in black in each plot.



Specificity of *Cre* x *R26-CAG-LSL-Sun1-sfGFP-myc*

Cre **(GFP)**		**anti-GAD67**	**anti-PV**	**anti-VIP**
Camk2a-Cre (n=2)	% GFP+ that are + for	1.0%	1.9%	1.0%
	% marker+ that are + for GFP	5.7%	7.6%	7.5%
PV-Cre (n=2)	% GFP+ that are + for		96.2%	0%
VIP-Cre (n=2)	% GFP+ that are + for		0%	97.4%

Figure 3. Widespread Differences in Gene Expression and DNA Methylation Among Neuron Subtypes

(A) Browser representation of RNA-seq read density and DNA methylation in CG and non-CG contexts (mCG, mCH) at two genes. *Slc6a1* (GAT-1, left) is expressed primarily in inhibitory neurons. *Lhx6* (right) is PV neuron-specific. Methylated CG (green) and CH (blue) positions are marked by upward (plus strand) and downward (minus strand) ticks. The height of each tick represents the % methylation, ranging from 0 to 100%. NeuN+ and Ctx (cortex) adult mouse methylomes are from Lister et al., 2013. R1, replicate 1; R2, replicate 2.

(B) Pairwise comparisons of protein-coding gene expression measured by RNA-seq across cell types (left three panels) or between replicates (right panel). The most differentially expressed genes (>5-fold change) are shown as colored points, and selected cell type-specific genes are labeled. r, Pearson correlation of log(TPM+0.1); TPM, transcripts per million.

(C) Percentage of cytosines in the CG and CH contexts that are methylated in each cell type.

(D) Percentage of all cytosines that are methylated. The number in each bar indicates the percentage of all methylated cytosines that occur in the CH context.

(E) Median ± 1 SEM of % mCH within and surrounding gene bodies, showing an inverse correlation between expression and DNA methylation at differentially expressed genes identified from our RNA-seq data (>5 fold-change for one cell type relative to both of the other cell types). TSS, transcription start site; TES, transcription end site.

(F) Pairwise comparisons of gene body % mCH across cell types (left three panels) or between replicates (right panel). Colored dots correspond to the same genes shown in Figure 3B.

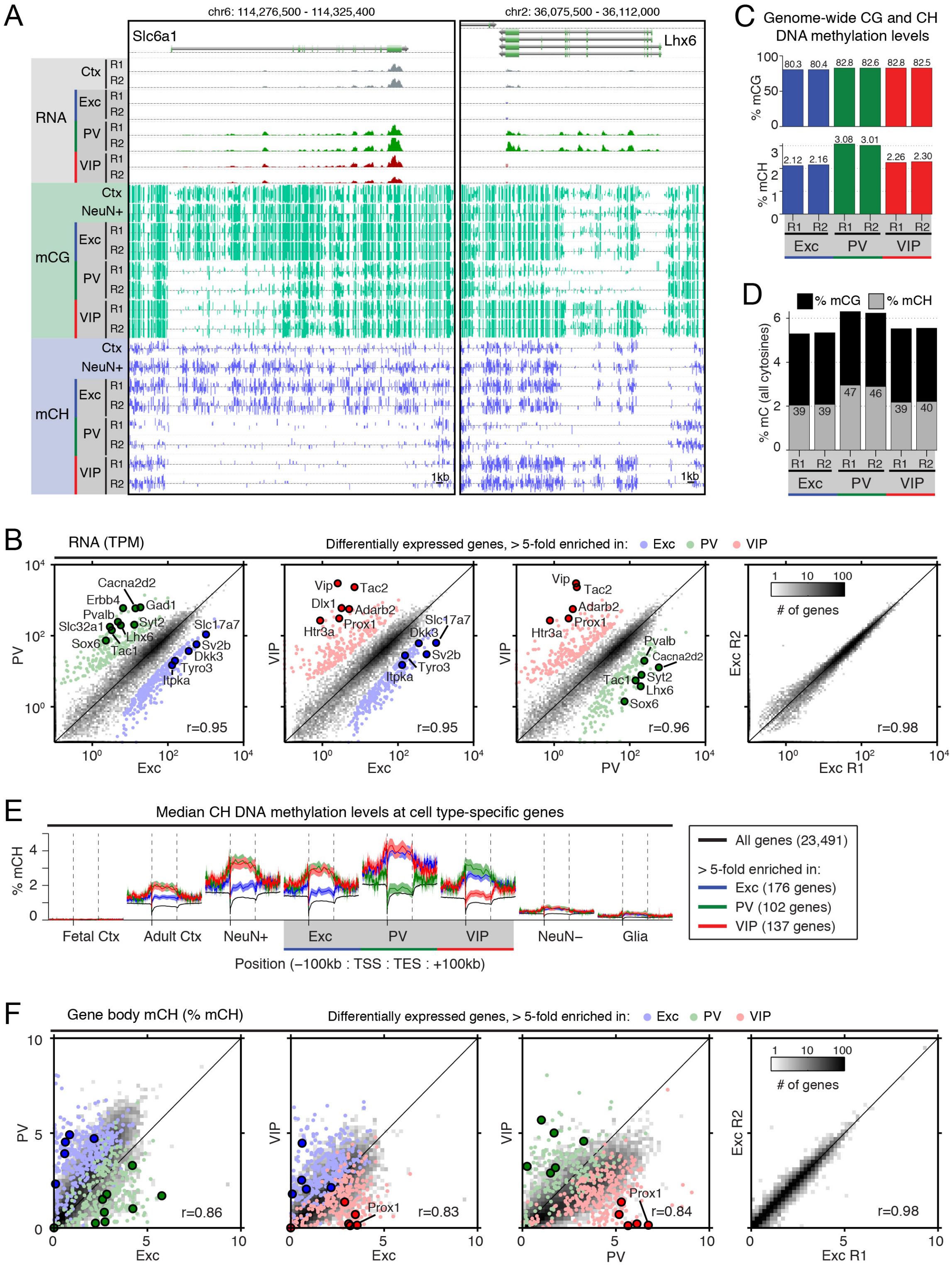

A
chr6: 114,276,500 - 114,325,400
chr2: 36,075,500 - 36,112,000
Slc6a1
Lhx6
RNA
mCG
mCH
Ctx
NeuN+
Exc
PV
VIP
R1
R2
1kb
C
Genome-wide CG and CH DNA methylation levels
% mCG
% mCH
100
50
0
80.3
80.4
82.8
82.6
82.8
82.5
3
2
1
0
2.12
2.16
3.08
3.01
2.26
2.30
R1 R2 R1 R2 R1 R2
Exc PV VIP
D
% mCG
% mCH
% mC (all cytosines)
6
4
2
0
39
39
47
46
39
40
R1 R2 R1 R2 R1 R2
Exc PV VIP
B
RNA (TPM)
Differentially expressed genes, > 5-fold enriched in: Exc PV VIP
10^4
10^2
10^0
PV
Exc
VIP
Exc R2
Exc R1
Cacna2d2
Erbb4
Gad1
Pvalb
Syt2
Slc32a1
Lhx6
Sox6
Tac1
Slc17a7
Sv2b
Dkk3
Tyro3
Itpka
r=0.95
Vip
Tac2
Dlx1
Adarb2
Prox1
Htr3a
r=0.95
r=0.96
r=0.98
1
10
100
# of genes
E
Median CH DNA methylation levels at cell type-specific genes
% mCH
4
2
0
Fetal Ctx
Adult Ctx
NeuN+
Exc
PV
VIP
NeuN−
Glia
Position (−100kb : TSS : TES : +100kb)
All genes (23,491)
> 5-fold enriched in:
Exc (176 genes)
PV (102 genes)
VIP (137 genes)
F
Gene body mCH (% mCH)
Differentially expressed genes, > 5-fold enriched in: Exc PV VIP
10
5
0
r=0.86
r=0.83
r=0.84
r=0.98
Prox1

Figure 4. Gene Expression and DNA Methylation Analysis

(A) Pairwise comparisons of gene expression levels between replicates in PV (left) and VIP (right) neurons. r, Pearson correlation of log(TPM+0.1). TPM, transcripts per million.

(B) Scatterplot showing high correlation between gene expression of INTACT-purified (*PV-Cre; R26-CAG-LSL-Sun1-sfGFP-myc*) and manually-sorted (G42 transgenic; Okaty et al., 2009) PV neurons. Selected cell type-specific genes are labeled (blue, Exc; green, PV; red, VIP) as well as candidate PV-enriched genes (black) tested by *in situ* hybridization. r, Spearman correlation.

(C) Double fluorescent ISH showing correct co-localization for nine genes predicted to be enriched in excitatory (left) or PV (right) neurons. *Slc17a7*, *Pvalb*, and *Vip* mark excitatory, PV, and VIP neurons, respectively. A 10th probe (*Zfp536*) did not co-localize with *Slc17a7*, *Pvalb*, or *Vip* at our level of detection (data not shown), and probe labeling was presumably in oligodendrocytes (Dugas et al., 2006).

(D) Barplot showing % mC for each non-CG methylation trinucleotide context.

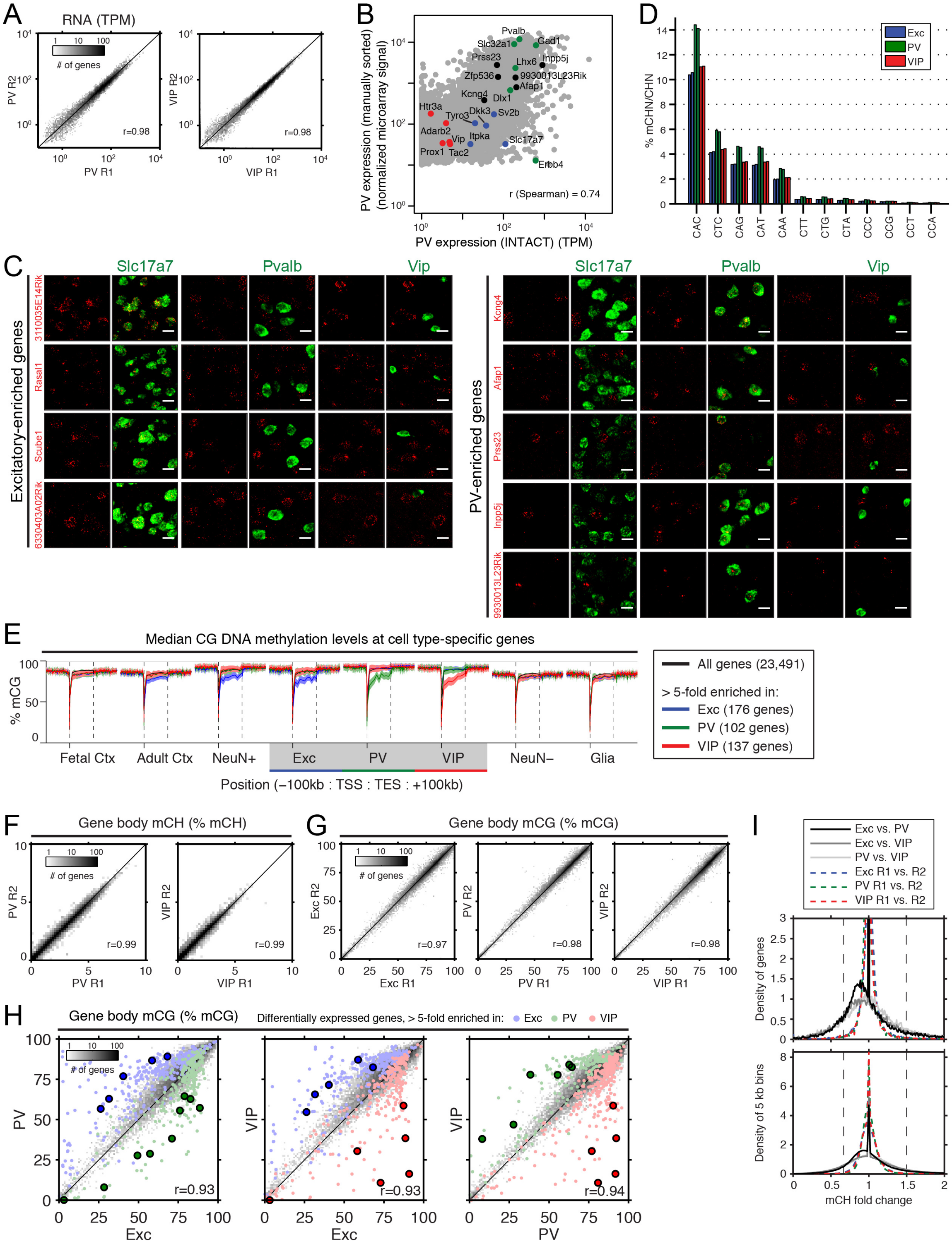
(E) Median ± 1 SEM of % mCG within and surrounding gene bodies, showing an inverse correlation between expression and DNA methylation at differentially expressed genes determined from our RNA-seq data (>5 fold-change for one cell type relative to both of the other cell types). TSS, transcription start site; TES, transcription end site.

(F) Pairwise comparisons of gene body % mCH between replicates in PV (left) and VIP (right) neurons.

(G) Pairwise comparisons of gene body % mCG between replicates in excitatory (left), PV (middle), and VIP (right).

(H) Pairwise comparisons of gene body % mCG across cell types. Colored dots correspond to the same genes shown in Figure 3B.

(I) Density plots showing ratios of CH methylation in gene bodies (top) and in 5 kb genomic bins (bottom) across cell types and between replicates. Each distribution was normalized by the median ratio. Dotted lines are at 0.67 and 1.5.



A
RNA (TPM)
# of genes
1 10 100
PV R2
PV R1
r=0.98
VIP R2
VIP R1
r=0.98
B
PV expression (manually sorted) (normalized microarray signal)
PV expression (INTACT) (TPM)
r (Spearman) = 0.74
Pvalb
Slc32a1
Gad1
Prss23
Lhx6
Inpp5j
Zfp536
9930013L23Rik
Afap1
Kcng4
Dlx1
Htr3a
Dkk3
Tyro3
Sv2b
Adarb2
Vip
Itpka
Slc17a7
Prox1
Tac2
Erbb4
D
% mCHN/CHN
Exc
PV
VIP
CAC
CTC
CAG
CAT
CAA
CTT
CTG
CTA
CCC
CCG
CCT
CCA
C
Slc17a7
Pvalb
Vip
Excitatory-enriched genes
3110035E14Rik
Rasal1
Scube1
6330403A02Rik
PV-enriched genes
Kcng4
Afap1
Prss23
Inpp5j
9930013L23Rik
E
Median CG DNA methylation levels at cell type-specific genes
% mCG
Fetal Ctx
Adult Ctx
NeuN+
Exc
PV
VIP
NeuN–
Glia
Position (–100kb : TSS : TES : +100kb)
All genes (23,491)
> 5-fold enriched in:
Exc (176 genes)
PV (102 genes)
VIP (137 genes)
F
Gene body mCH (% mCH)
# of genes
PV R2
PV R1
r=0.99
VIP R2
VIP R1
r=0.99
G
Gene body mCG (% mCG)
Exc R2
Exc R1
r=0.97
PV R2
PV R1
r=0.98
VIP R2
VIP R1
r=0.98
H
Gene body mCG (% mCG)
Differentially expressed genes, > 5-fold enriched in: Exc PV VIP
# of genes
PV
Exc
r=0.93
VIP
Exc
r=0.93
VIP
PV
r=0.94
I
Exc vs. PV
Exc vs. VIP
PV vs. VIP
Exc R1 vs. R2
PV R1 vs. R2
VIP R1 vs. R2
Density of genes
Density of 5 kb bins
mCH fold change

Figure 5. Epigenomic Marks are Coordinated and Highly Cell Type-Specific

(A) Examples of intergenic regulatory elements marked by accessible chromatin (peaks in ATAC-seq read density, upper tracks) and low levels of DNA methylation (hypo-DMRs and UMRs+LMRs, lower tracks) at an intergenic region ~53 kb upstream of *Snap25* (both the nearest gene and the nearest TSS). Locations of ATAC-seq peaks, hypo-DMRs, and UMRs+LMRs are shown below the corresponding raw data. R1, replicate 1; R2, replicate 2.

(B) Area-proportional Venn diagram showing the numbers of all cell type-specific and shared ATAC-seq peaks across excitatory, PV, and VIP neurons (top). Area-proportional Venn diagrams showing that a greater fraction of promoter-associated peaks (within 2.5 kb of a TSS) are shared compared to distal peaks (>20 kb from a TSS), which are predominantly cell type-specific (bottom).

(C) Browser representation of regulatory elements around *trkC*/*Ntrk3* marked by histone modifications in excitatory neurons, DNaseI hypersensitivity in whole cerebrum (from ENCODE), and peaks in ATAC-seq read density in excitatory, PV, and VIP neurons. For ATAC-seq, greater spatial resolution is achieved by using reads <100 bp in length (tracks labeled <100).

(D) Area-proportional Venn diagram showing the numbers of DMRs identified to be hypo-methylated in excitatory, PV, and/or VIP neurons in a statistical comparison of CG methylation levels across five cell types. Two of these cell types, fetal cortex and glia, are not shown in the diagram.

(E) Heatmap showing % mCG plotted in 3 kb windows centered at DMRs hypo-methylated in one or two cell types (panel 1). At the same genomic regions, the following additional features were plotted: % mCH (panel 2), chromatin accessibility (ATAC-seq reads) (panel 3), and histone modification ChIP-seq reads in excitatory neurons (panel 4). The number of DMRs in each category is shown in parentheses.

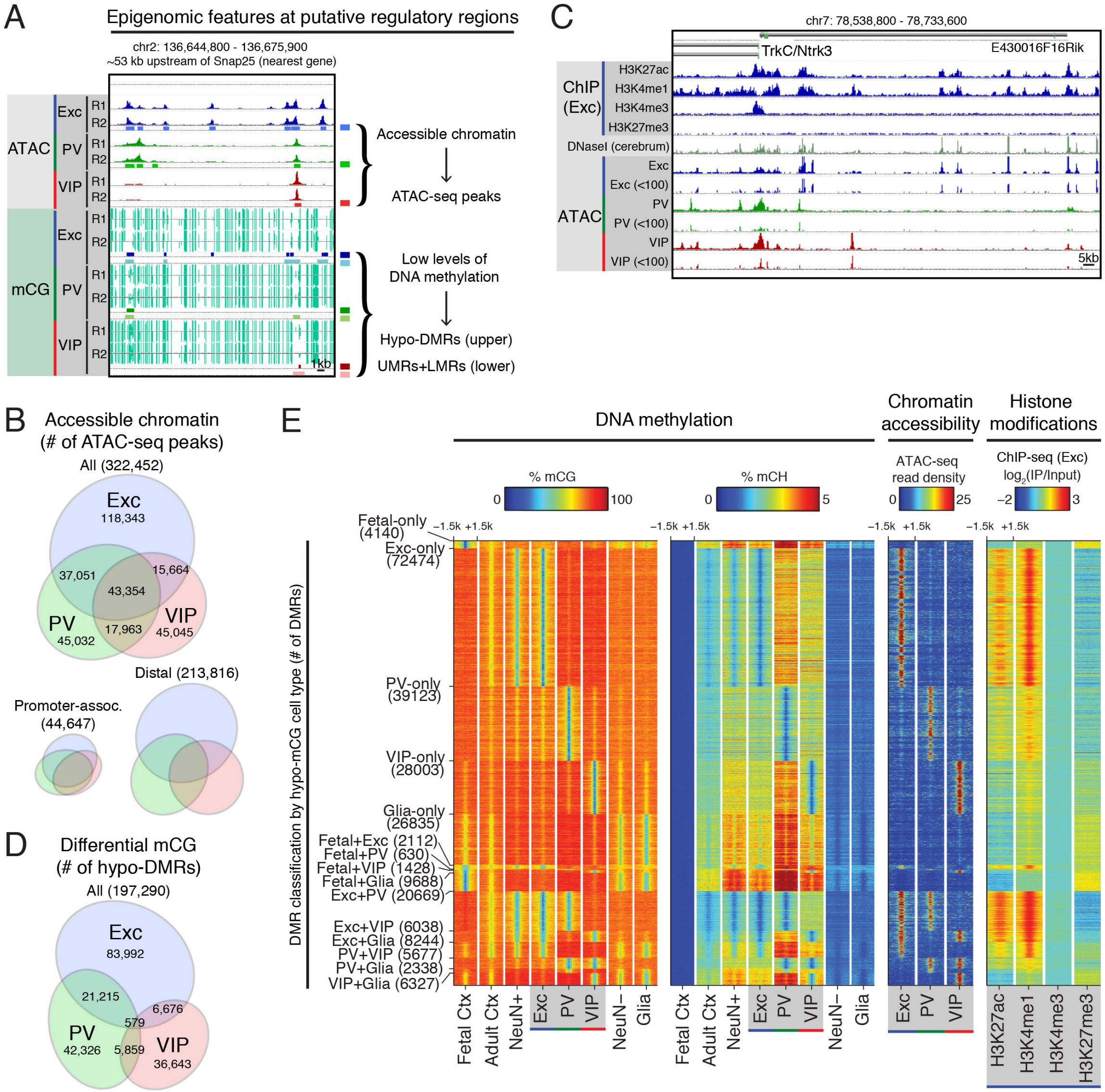

A
Epigenomic features at putative regulatory regions
chr2: 136,644,800 - 136,675,900
~53 kb upstream of Snap25 (nearest gene)
ATAC
Exc
PV
VIP
R1
R2
mCG
1kb
Accessible chromatin
ATAC-seq peaks
Low levels of DNA methylation
Hypo-DMRs (upper)
UMRs+LMRs (lower)
B
Accessible chromatin (# of ATAC-seq peaks)
All (322,452)
Exc 118,343
37,051
15,664
43,354
PV 45,032
17,963
VIP 45,045
Distal (213,816)
Promoter-assoc. (44,647)
C
chr7: 78,538,800 - 78,733,600
TrkC/Ntrk3
E430016F16Rik
ChIP (Exc)
H3K27ac
H3K4me1
H3K4me3
H3K27me3
DNaseI (cerebrum)
ATAC
Exc
Exc (<100)
PV
PV (<100)
VIP
VIP (<100)
5kb
D
Differential mCG (# of hypo-DMRs)
All (197,290)
Exc 83,992
21,215
6,676
579
PV 42,326
5,859
VIP 36,643
E
DNA methylation
Chromatin accessibility
Histone modifications
% mCG
0
100
% mCH
0
5
ATAC-seq read density
0
25
ChIP-seq (Exc) $\log_2$(IP/Input)
−2
3
−1.5k +1.5k
DMR classification by hypo-mCG cell type (# of DMRs)
Fetal-only (4140)
Exc-only (72474)
PV-only (39123)
VIP-only (28003)
Glia-only (26835)
Fetal+Exc (2112)
Fetal+PV (630)
Fetal+VIP (1428)
Fetal+Glia (9688)
Exc+PV (20669)
Exc+VIP (6038)
Exc+Glia (8244)
PV+VIP (5677)
PV+Glia (2338)
VIP+Glia (6327)
Fetal Ctx
Adult Ctx
NeuN+
Exc
PV
VIP
NeuN−
Glia
H3K27ac
H3K4me1
H3K4me3
H3K27me3

Figure 6. Correlations Across Epigenomic Marks and Relevance of Neuron Subtype-Specific Hypo-Methylation to Induced Neuronal Activity

(A) Examples of regulatory elements marked by accessible chromatin (peaks in ATAC-seq read density, upper tracks) and low levels of DNA methylation (hypo-DMRs and UMRs+LMRs, lower tracks) near *Ngf*. Locations of ATAC-seq peaks, hypo-DMRs, and UMRs+LMRs are shown below the corresponding raw reads. R1, replicate 1; R2, replicate 2.

(B) Barplot showing that the majority of binding sites of six activity-dependent TFs in KCl-depolarized cortical cultures (Kim et al., 2010; Malik et al., 2014) overlap with excitatory neuron UMRs+LMRs (left). Binding sites for FOS, FOSB, JUNB, and NPAS4 also overlap extensively with excitatory-specific hypo-DMRs. The number of total ChIP-seq peaks for each TF is shown in parentheses. Barplot showing the enrichment and depletion of each hypo-DMR category overlapping TF ChIP-seq peaks (right). CREB and SRF were excluded since their enrichments and depletions were insignificant at q<1E-5.

(C) At the same regions as in Figure 5E (i.e., 3 kb windows centered at DMRs hypo-methylated in one or two cell types), heatmap showing TF ChIP-seq reads from unstimulated (un) and depolarized (KCl) cortical cultures (TF ChIP-seq from Kim et al., 2010; Malik et al., 2014).

(D) Scatterplots showing high correlation between mCG and mCH at DMRs. mCG and mCH levels in each DMR were normalized by the mean mCG and mCH in that DMR across the three cell types. r, Pearson correlation.

(E) A matrix showing pairwise Pearson correlations of H3K27ac, H3K4me1, and sub-nucleosomal (<100 bp) ATAC-seq reads at enhancers. H3K4me1 and H3K27ac signals are generally well-correlated at a global level; however, individual enhancers can be poised (H3K4me1+; 42,540 enhancers) or active (H3K4me1+/H3K27ac+; 48,781 enhancers). ATAC-seq signal is also correlated, albeit to a lesser degree (r ~0.5), with both H3K4me1 and H3K27ac signal at enhancers. R1, replicate 1; R2, replicate 2.

(F) In each cell type, the majority of hypo-DMRs are a subset of UMRs+LMRs. The majority of ATAC-seq peaks and UMRs+LMRs overlap. For excitatory neurons, approximately half of ATAC-seq peaks, hypo-DMRs, and UMRs+LMRs overlap with enhancers identified using histone modifications. Although these DNA methylation and chromatin features are overlapping, they are not synonymous. Part of the difference may arise from statistical thresholds set in the identification of each region; however, each type of dataset also provides non-redundant and complementary information that depend on the genomic context. The numbers in parentheses indicate the total number of these features identified in each cell type.

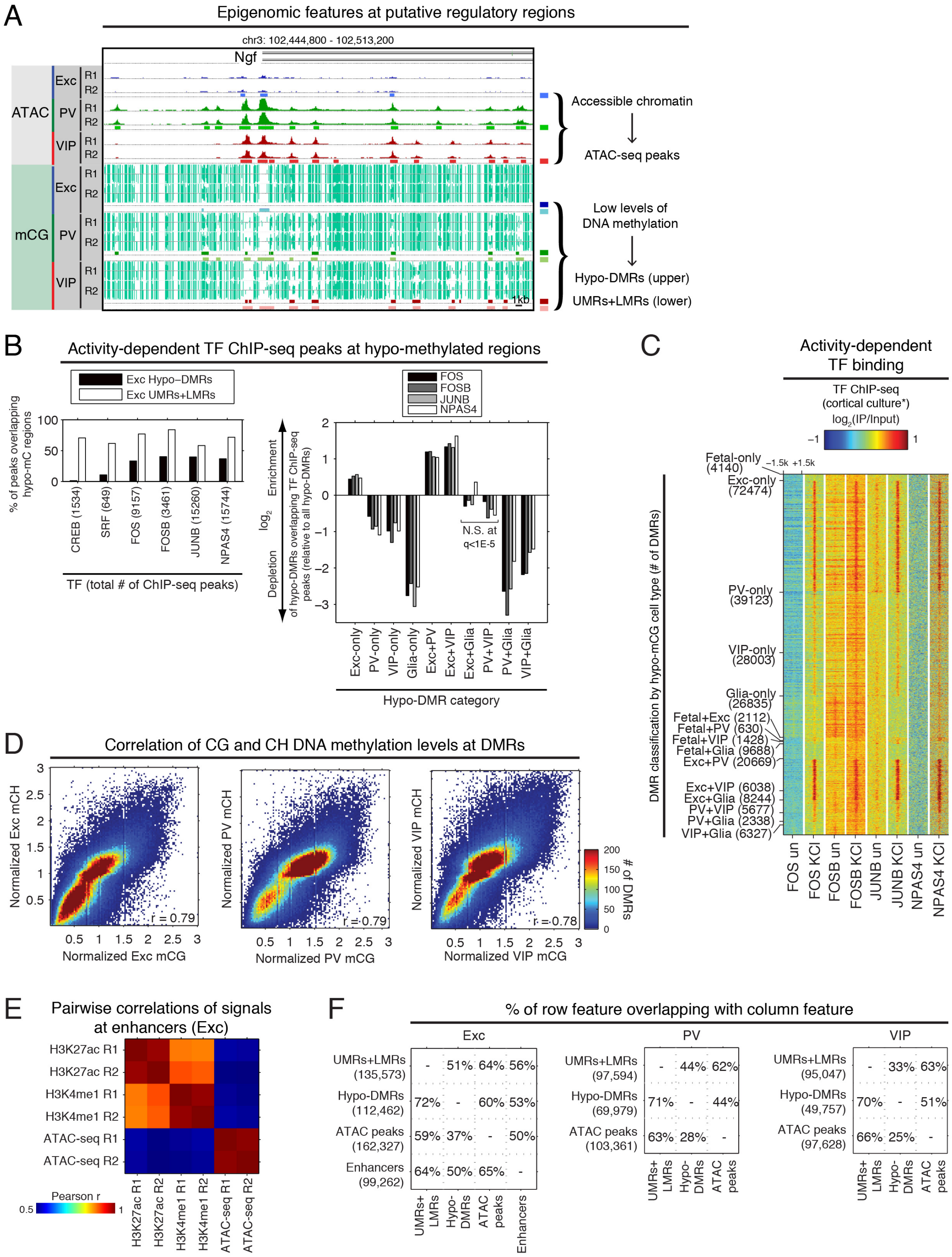

A
Epigenomic features at putative regulatory regions
chr3: 102,444,800 - 102,513,200
Ngf
ATAC
mCG
Exc
PV
VIP
R1
R2
1kb
Accessible chromatin
ATAC-seq peaks
Low levels of DNA methylation
Hypo-DMRs (upper)
UMRs+LMRs (lower)
B
Activity-dependent TF ChIP-seq peaks at hypo-methylated regions
Exc Hypo−DMRs
Exc UMRs+LMRs
% of peaks overlapping hypo-mC regions
100
50
0
CREB (1534)
SRF (649)
FOS (9157)
FOSB (3461)
JUNB (15260)
NPAS4 (15744)
TF (total # of ChIP-seq peaks)
FOS
FOSB
JUNB
NPAS4
Enrichment
Depletion
log2 of hypo-DMRs overlapping TF ChIP-seq peaks (relative to all hypo-DMRs)
2
1
0
−1
−2
−3
N.S. at q<1E-5
Exc-only
PV-only
VIP-only
Glia-only
Exc+PV
Exc+VIP
Exc+Glia
PV+VIP
PV+Glia
VIP+Glia
Hypo-DMR category
C
Activity-dependent TF binding
TF ChIP-seq (cortical culture*)
log2(IP/Input)
−1
1
−1.5k
+1.5k
DMR classification by hypo-mCG cell type (# of DMRs)
Fetal-only (4140)
Exc-only (72474)
PV-only (39123)
VIP-only (28003)
Glia-only (26835)
Fetal+Exc (2112)
Fetal+PV (630)
Fetal+VIP (1428)
Fetal+Glia (9688)
Exc+PV (20669)
Exc+VIP (6038)
Exc+Glia (8244)
PV+VIP (5677)
PV+Glia (2338)
VIP+Glia (6327)
FOS un
FOS KCl
FOSB un
FOSB KCl
JUNB un
JUNB KCl
NPAS4 un
NPAS4 KCl
D
Correlation of CG and CH DNA methylation levels at DMRs
Normalized Exc mCH
Normalized PV mCH
Normalized VIP mCH
Normalized Exc mCG
Normalized PV mCG
Normalized VIP mCG
0.5
1
1.5
2
2.5
3
r = 0.79
r = 0.79
r = 0.78
# of DMRs
200
150
100
50
0
E
Pairwise correlations of signals at enhancers (Exc)
H3K27ac R1
H3K27ac R2
H3K4me1 R1
H3K4me1 R2
ATAC-seq R1
ATAC-seq R2
Pearson r
0.5
1
F
% of row feature overlapping with column feature
Exc
UMRs+LMRs (135,573) - 51% 64% 56%
Hypo-DMRs (112,462) 72% - 60% 53%
ATAC peaks (162,327) 59% 37% - 50%
Enhancers (99,262) 64% 50% 65% -
UMRs+ LMRs
Hypo- DMRs
ATAC peaks
Enhancers
PV
UMRs+LMRs (97,594) - 44% 62%
Hypo-DMRs (69,979) 71% - 44%
ATAC peaks (103,361) 63% 28% -
UMRs+ LMRs
Hypo- DMRs
ATAC peaks
VIP
UMRs+LMRs (95,047) - 33% 63%
Hypo-DMRs (49,757) 70% - 51%
ATAC peaks (97,628) 66% 25% -
UMRs+ LMRs
Hypo- DMRs
ATAC peaks

Figure 7. Relationships Across Cell Types and Development Via Epigenomic Marks

(A-B) Matrices showing pairwise Pearson correlations for % mCG (A) and ATAC-seq read densities (B) at ATAC-seq peaks. Dendrograms show hierarchical clustering using complete linkage and 1-Pearson correlation as the metric.

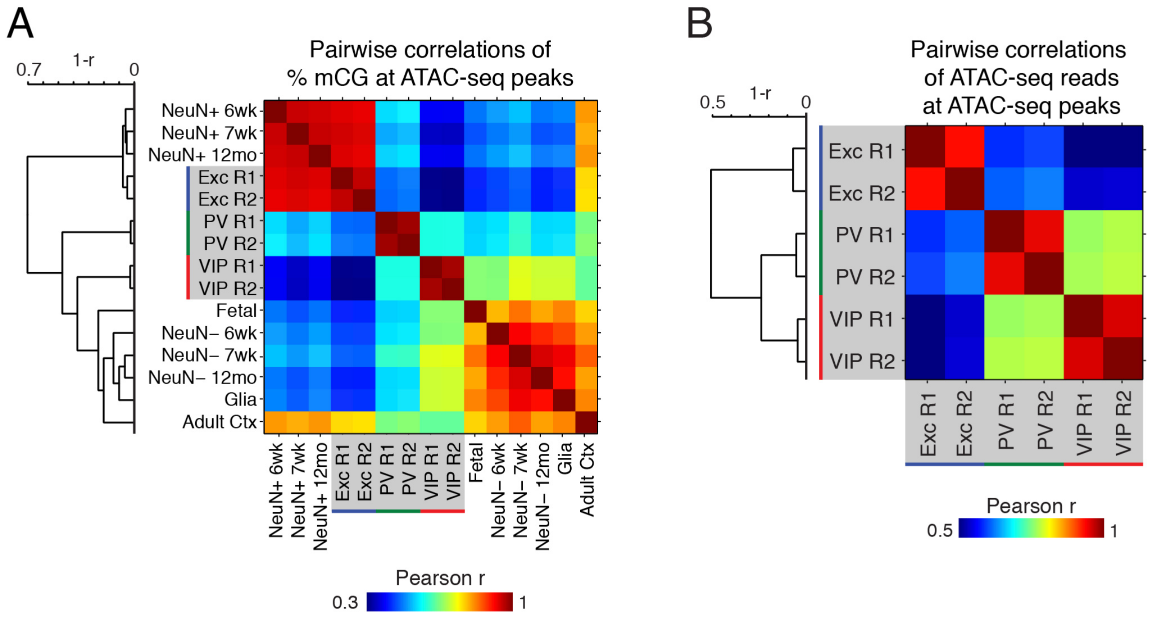
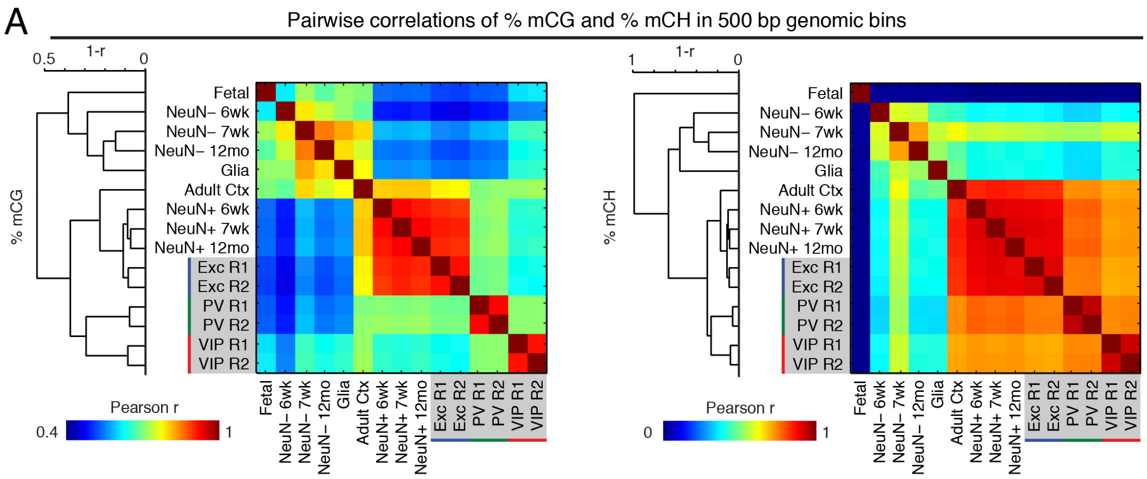


Figure 8. Epigenomic Correlations Across Cell Types and Development

(A) Matrices showing pairwise Pearson correlations for % mCG (left) and for % mCH (right) in 500 bp genomic bins across all autosomes. Dendrograms show hierarchical clustering using complete linkage and 1-Pearson correlation as the metric.

(B) A matrix showing pairwise Pearson correlations for % mCH at ATAC-seq peaks. The dendrogram shows hierarchical clustering using complete linkage and 1-Pearson correlation as the metric.



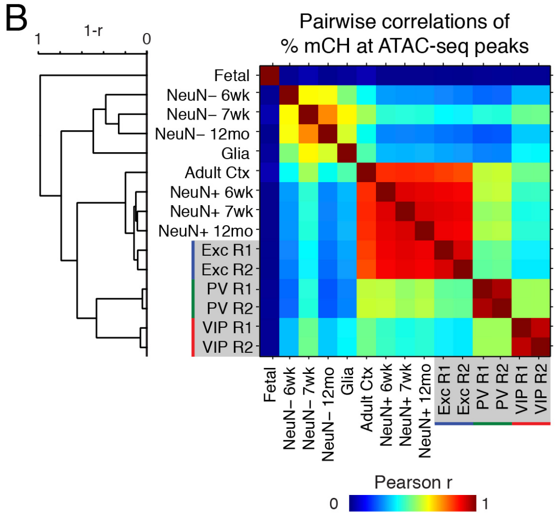


Figure 9. Neuronal Subtypes are Associated with Distinct Patterns of TF Binding

(A) The average density of ATAC-seq read endpoints (Tn5 transposase insertions) within ±100 bp relative to the estimated locations of footprints for four example TFs, showing characteristic footprint structures. Each footprint profile is normalized by the maximum over the profiled region. Inset: position weight matrix showing conserved sequence motifs at the footprint center.

(B) Heatmaps showing the enrichment (red) and depletion (blue) of footprints in cell type-specific ATAC-seq peaks (left) or motifs in hypo-DMRs (middle). The relative TF expression level across excitatory, PV, and VIP neurons is also shown (right). Selected TFs are labeled.

(C) Schematic for assessing TF-TF interactions by detecting footprints of one TF (FP A) in a 20 kb window around the TSS of a second TF (TF B); footprints located farther away (FP C) are not predicted to interact.

(D) Networks of TF interactions predicted by the method shown in (C) using cell type-specific and pan-neuronal footprints.

(E) Heatmaps showing the average density of cell type-specific and pan-neuronal footprints within a TSS±100 kb window for each category of genes.

(F) Barplot showing the average % of base pairs within a TSS±10 kb window that overlaps each ATAC-seq peak category, for each category of genes (left). Heatmap showing an enrichment of cell type-specific peaks at both cell type-specific and pan-neuronal genes (right). Pan-neuronal genes are from Hobert et al., 2010; q from 1-sided Wilcoxon rank sum test with Benjamini-Hochberg FDR correction.

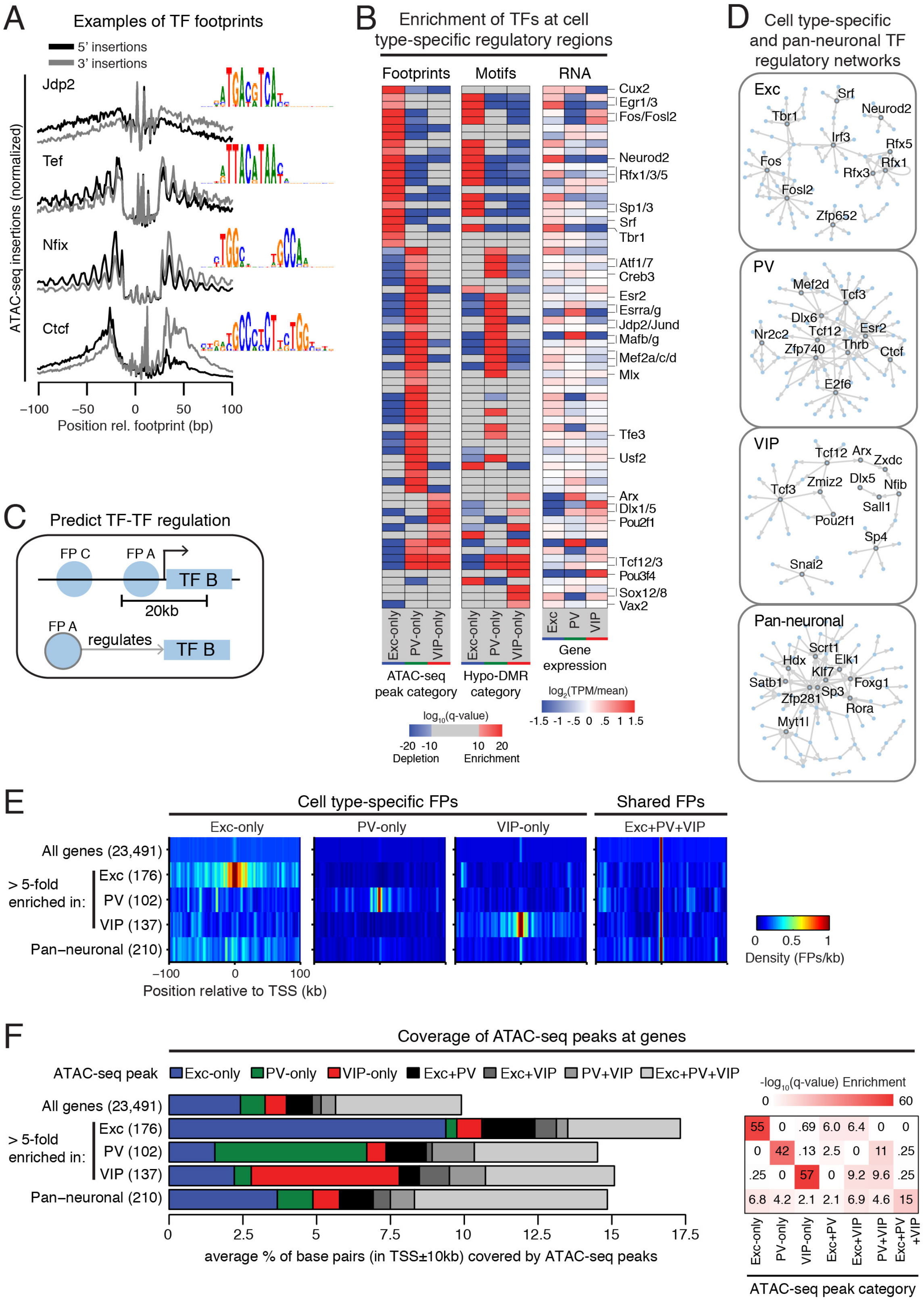

A
Examples of TF footprints
5' insertions
3' insertions
ATAC-seq insertions (normalized)
Jdp2
Tef
Nfix
Ctcf
−100
−50
0
50
100
Position rel. footprint (bp)
B
Enrichment of TFs at cell type-specific regulatory regions
Footprints
Motifs
RNA
Cux2
Egr1/3
Fos/Fosl2
Neurod2
Rfx1/3/5
Sp1/3
Srf
Tbr1
Atf1/7
Creb3
Esr2
Esrra/g
Jdp2/Jund
Mafb/g
Mef2a/c/d
Mlx
Tfe3
Usf2
Arx
Dlx1/5
Pou2f1
Tcf12/3
Pou3f4
Sox12/8
Vax2
Exc-only
PV-only
VIP-only
Exc-only
PV-only
VIP-only
Exc
PV
VIP
ATAC-seq peak category
Hypo-DMR category
Gene expression
log$_{10}$(q-value)
-20 -10
10 20
Depletion
Enrichment
log$_2$(TPM/mean)
-1.5 -1 -.5 0 .5 1 1.5
C
Predict TF-TF regulation
FP C
FP A
TF B
20kb
FP A
regulates
TF B
D
Cell type-specific and pan-neuronal TF regulatory networks
Exc
Srf
Neurod2
Tbr1
Irf3
Rfx5
Fos
Rfx1
Rfx3
Fosl2
Zfp652
PV
Mef2d
Tcf3
Dlx6
Nr2c2
Tcf12
Esr2
Zfp740
Thrb
Ctcf
E2f6
VIP
Tcf12
Arx
Zxdc
Zmiz2
Dlx5
Nfib
Tcf3
Sall1
Pou2f1
Sp4
Snai2
Pan-neuronal
Scrt1
Hdx
Elk1
Klf7
Satb1
Foxg1
Sp3
Zfp281
Rora
Myt1l
E
Cell type-specific FPs
Shared FPs
Exc-only
PV-only
VIP-only
Exc+PV+VIP
All genes (23,491)
> 5-fold enriched in:
Exc (176)
PV (102)
VIP (137)
Pan−neuronal (210)
−100
0
100
Position relative to TSS (kb)
0
0.5
1
Density (FPs/kb)
F
Coverage of ATAC-seq peaks at genes
ATAC-seq peak
Exc-only
PV-only
VIP-only
Exc+PV
Exc+VIP
PV+VIP
Exc+PV+VIP
All genes (23,491)
> 5-fold enriched in:
Exc (176)
PV (102)
VIP (137)
Pan−neuronal (210)
0
2.5
5
7.5
10
12.5
15
17.5
average % of base pairs (in TSS±10kb) covered by ATAC-seq peaks
-log$_{10}$(q-value) Enrichment
0
60
55 0 .69 6.0 6.4 0 0
0 42 .13 2.5 0 11 .25
.25 0 57 0 9.2 9.6 .25
6.8 4.2 2.1 2.1 6.9 4.6 15
Exc-only
PV-only
VIP-only
Exc+PV
Exc+VIP
PV+VIP
Exc+PV +VIP
ATAC-seq peak category

Figure 10. Analysis of Putative TF Binding at Neuronal Regulatory Regions

(A) Scatterplots showing, for expressed TFs (TPM≥30), % mCG (top), % mCH (bottom left), and % mCA (bottom right) around regions that are footprinted in one cell type (y-axis) versus regions that are not footprinted in that cell type, but are footprinted in a different cell type (x-axis). Most TF footprints lie in regions of lower DNA methylation, relative to the methylation levels found in cell types without footprints for the same regions. Exceptions include CTCF and ZFP410.

(B) Heatmap showing TF motif enrichments (left) and gene expression (right) for all categories of DMRs that are hypo-methylated in one or two cell types across excitatory, PV, and VIP neurons as well as glia and fetal cortex. Boxes indicate TFs mentioned in the main text.

(C) Examples of pan-neuronal genes (from Hobert et al., 2010) surrounded by cell type-specific and pan-neuronal regions of increased chromatin accessibility, as determined by peaks of ATAC-seq read density. Arrows point to a subset of cell type-specific ATAC-seq peaks. R1, replicate 1; R2, replicate 2.

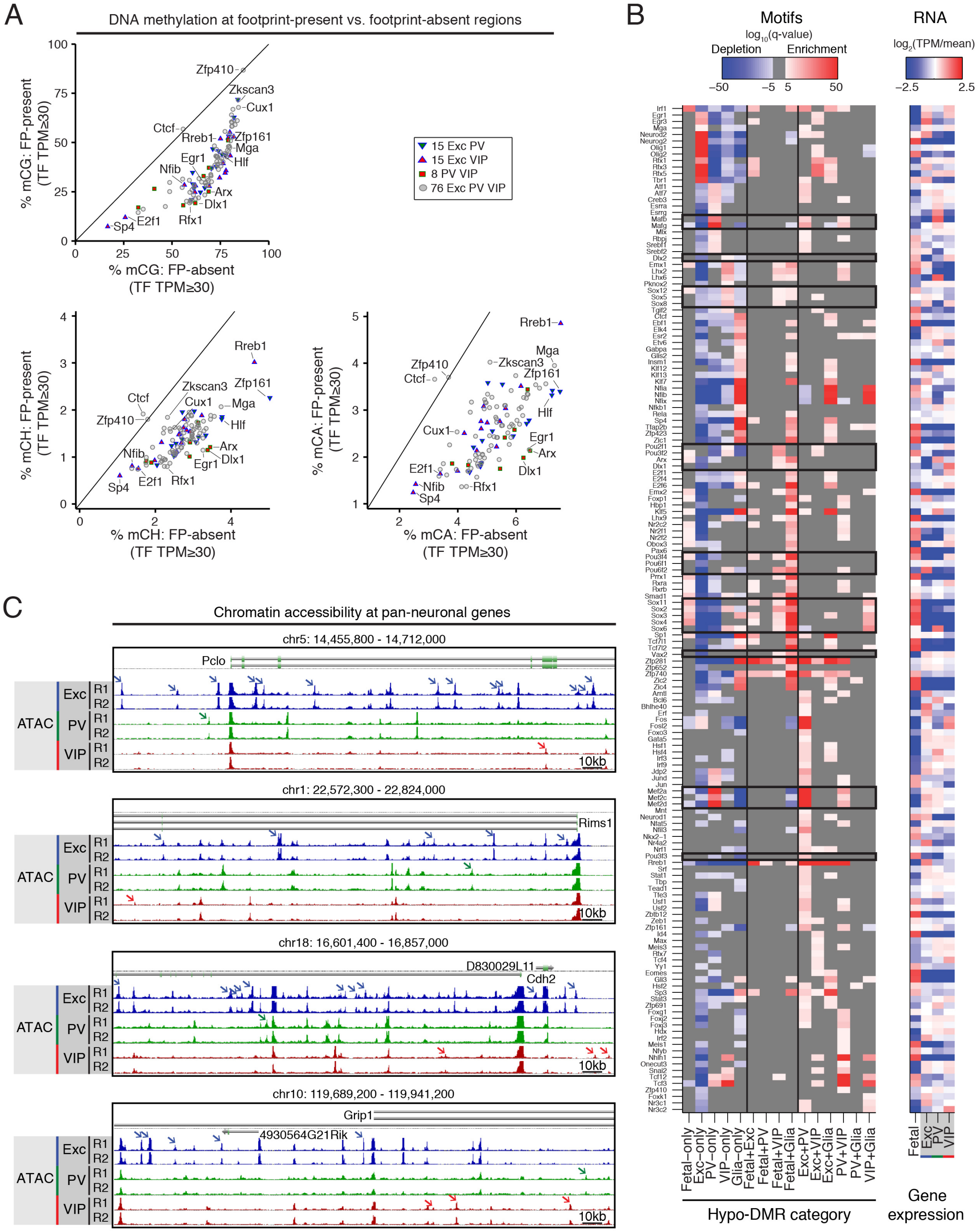

A
DNA methylation at footprint-present vs. footprint-absent regions
% mCG: FP-present (TF TPM≥30)
% mCG: FP-absent (TF TPM≥30)
Zfp410
Zkscan3
Cux1
Ctcf
Rreb1
Zfp161
Mga
Hlf
Egr1
Nfib
Arx
Dlx1
Rfx1
E2f1
Sp4
15 Exc PV
15 Exc VIP
8 PV VIP
76 Exc PV VIP
% mCH: FP-present (TF TPM≥30)
% mCH: FP-absent (TF TPM≥30)
% mCA: FP-present (TF TPM≥30)
% mCA: FP-absent (TF TPM≥30)
B
Motifs
log10(q-value)
Depletion
Enrichment
−50
−5
5
50
RNA
log2(TPM/mean)
−2.5
2.5
Fetal−only
Exc−only
PV−only
VIP−only
Glia−only
Fetal+Exc
Fetal+PV
Fetal+VIP
Fetal+Glia
Exc+PV
Exc+VIP
Exc+Glia
PV+VIP
PV+Glia
VIP+Glia
Hypo-DMR category
Fetal
Exc
PV
VIP
Gene expression
C
Chromatin accessibility at pan-neuronal genes
chr5: 14,455,800 - 14,712,000
Pclo
chr1: 22,572,300 - 22,824,000
Rims1
chr18: 16,601,400 - 16,857,000
D830029L11
Cdh2
chr10: 119,689,200 - 119,941,200
Grip1
4930564G21Rik
ATAC
Exc
PV
VIP
R1
R2
10kb

Figure 11. Integrative Analysis of DNA Methylation, Gene Expression, and Chromatin Features

(A) Spearman correlations of three epigenomic features (CG DNA methylation, CH DNA methylation, and ATAC-seq read density) with RNA expression level around the TSS of autosomal expressed (TPM>0.1) genes (left) and differentially expressed genes (right). Note that the signs of the correlations for mCG and mCH are negative (i.e., these features inversely correlate with gene expression).

(B-E) Protein-coding genes were clustered by k-means based on patterns of intragenic mCH. For each cluster (1-25), the following features are plotted: mCH level within each gene body and flanking 100 kb (B); mRNA abundance (C); enrichment or depletion for differentially expressed (DE) genes (D), and enrichment or depletion for cell type-specific and shared ATAC-seq peaks within ±10 kb of the TSS (E). mCH levels for each gene are normalized by the levels at distal flanking regions (50-100 kb upstream and downstream of the gene body). For clusters with cell type-specific hypo-methylation, an example gene or gene set is listed. TPM, transcripts per million; N.S., not significant (FET, q<0.01).

(F) mCH levels are higher in the nucleosomal linker region and lower in the nucleosome core. mCH levels are normalized by the level at flanking regions (1-2 kb upstream and downstream of the nucleosome center).

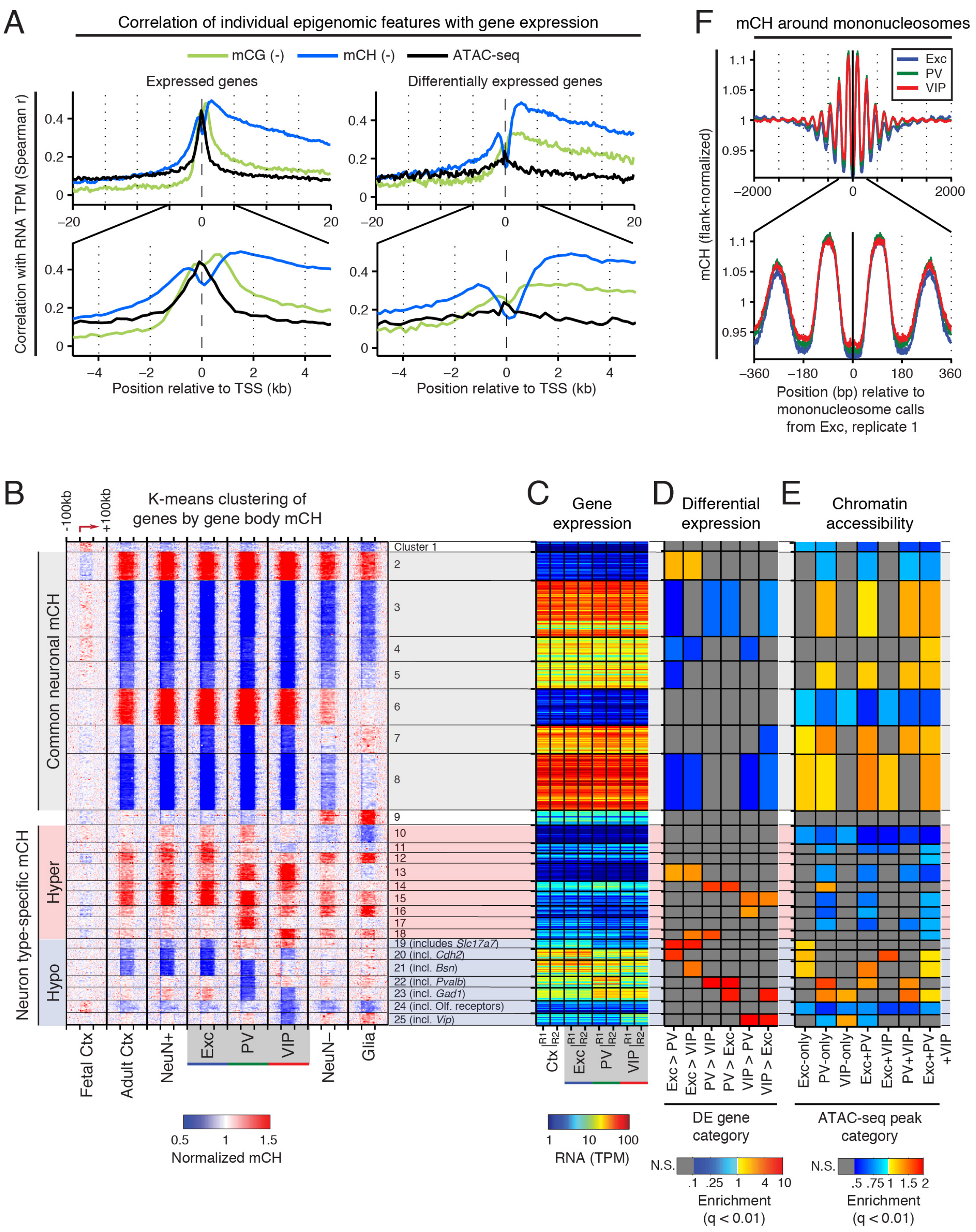

A
Correlation of individual epigenomic features with gene expression
mCG (-)
mCH (-)
ATAC-seq
Expressed genes
Differentially expressed genes
Correlation with RNA TPM (Spearman r)
0
0.2
0.4
–20
0
20
–4
–2
0
2
4
Position relative to TSS (kb)
F
mCH around mononucleosomes
Exc
PV
VIP
mCH (flank-normalized)
0.95
1
1.05
1.1
–2000
0
2000
–360
–180
0
180
360
Position (bp) relative to mononucleosome calls from Exc, replicate 1
B
K-means clustering of genes by gene body mCH
-100kb
+100kb
Common neuronal mCH
Neuron type-specific mCH
Hyper
Hypo
Cluster 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19 (includes *Slc17a7*)
20 (incl. *Cdh2*)
21 (incl. *Bsn*)
22 (incl. *Pvalb*)
23 (incl. *Gad1*)
24 (incl. Olf. receptors)
25 (incl. *Vip*)
Fetal Ctx
Adult Ctx
NeuN+
Exc
PV
VIP
NeuN–
Glia
0.5
1
1.5
Normalized mCH
C
Gene expression
Ctx
Exc
PV
VIP
R1
R2
1
10
100
RNA (TPM)
D
Differential expression
Exc > PV
Exc > VIP
PV > VIP
PV > Exc
VIP > PV
VIP > Exc
DE gene category
N.S.
.1
.25
1
4
10
Enrichment (q < 0.01)
E
Chromatin accessibility
Exc-only
PV-only
VIP-only
Exc+PV
Exc+VIP
PV+VIP
Exc+PV +VIP
ATAC-seq peak category
N.S.
.5
.75
1
1.5
2
Enrichment (q < 0.01)

Figure 12. Integrative Analysis of Epigenomic Features

(A) LASSO regression using the top 7 selected epigenomic features gives a Spearman correlation of 0.66, with intragenic non-CG methylation as the most informative feature. Epigenomic features used in the regression were mCG, mCH, ATAC-seq, and DMR density at different positions around genes. A.U., arbitrary units.

(B) Line plot showing that LASSO regression using more than ~7 features does not generate substantially higher correlations (top). The normalized fit coefficient for the 8 best features is shown as a function of the regularization parameter (bottom). The red square indicates 7 features.

(C) Choice of the number of clusters used for k-means clustering.

(D) Line plots showing lower mCG and ATAC-seq read density at the mononucleosome core.

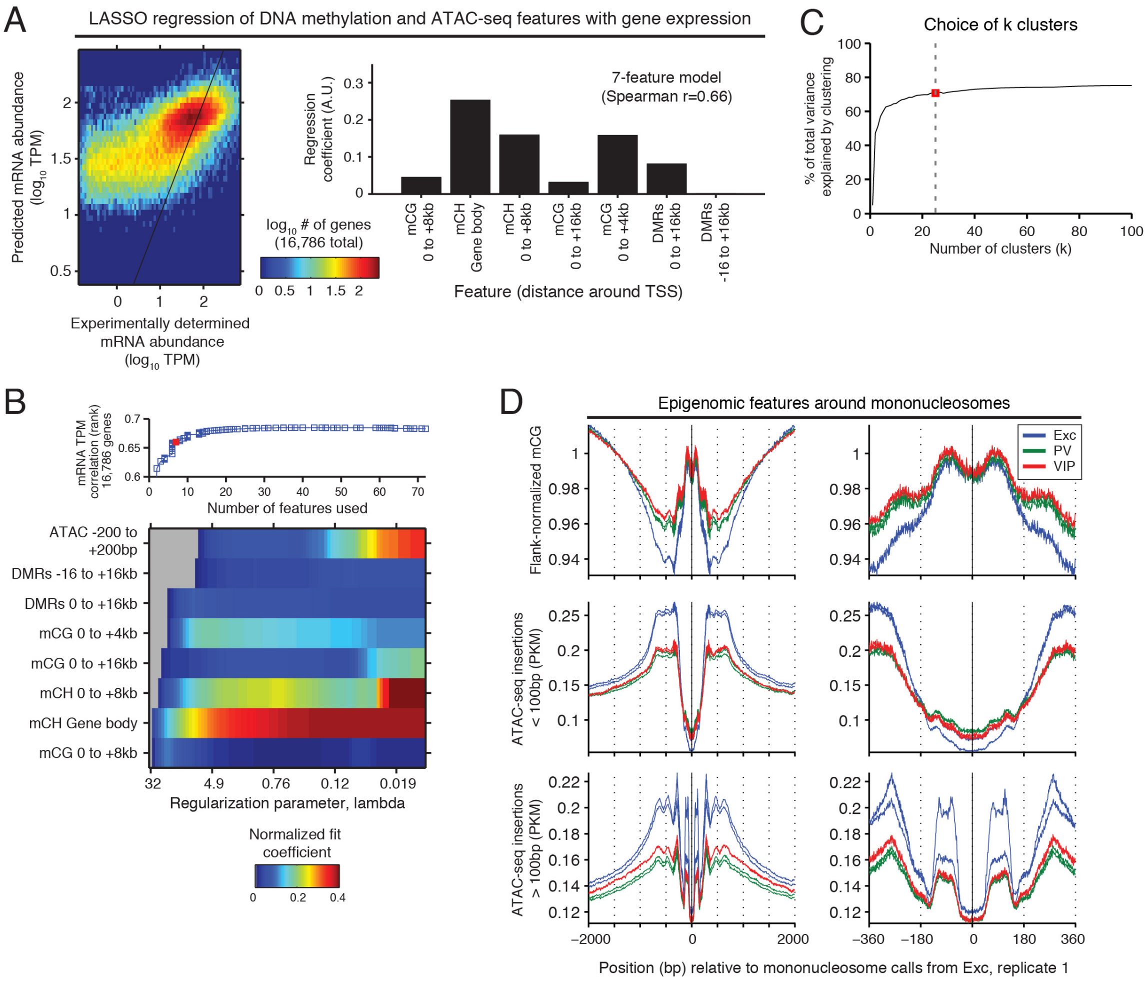

A
LASSO regression of DNA methylation and ATAC-seq features with gene expression
Predicted mRNA abundance ($\log_{10}$ TPM)
0.5
1
1.5
2
0
1
2
Experimentally determined mRNA abundance ($\log_{10}$ TPM)
$\log_{10}$ # of genes (16,786 total)
0 0.5 1 1.5 2
Regression coefficient (A.U.)
0
0.1
0.2
0.3
7-feature model (Spearman r=0.66)
mCG 0 to +8kb
mCH Gene body
mCH 0 to +8kb
mCG 0 to +16kb
mCG 0 to +4kb
DMRs 0 to +16kb
DMRs -16 to +16kb
Feature (distance around TSS)
C
Choice of k clusters
% of total variance explained by clustering
0
20
40
60
80
100
0
20
40
60
80
100
Number of clusters (k)
B
mRNA TPM correlation (rank) 16,786 genes
0.6
0.65
0.7
0
10
20
30
40
50
60
70
Number of features used
ATAC -200 to +200bp
DMRs -16 to +16kb
DMRs 0 to +16kb
mCG 0 to +4kb
mCG 0 to +16kb
mCH 0 to +8kb
mCH Gene body
mCG 0 to +8kb
32
4.9
0.76
0.12
0.019
Regularization parameter, lambda
Normalized fit coefficient
0
0.2
0.4
D
Epigenomic features around mononucleosomes
Exc
PV
VIP
Flank-normalized mCG
0.94
0.96
0.98
1
ATAC-seq insertions < 100bp (PKM)
0.1
0.15
0.2
0.25
ATAC-seq insertions > 100bp (PKM)
0.12
0.14
0.16
0.18
0.2
0.22
–2000
0
2000
–360
–180
0
180
360
Position (bp) relative to mononucleosome calls from Exc, replicate 1

Figure 13. Large Domains of Low Methylation Link to Gene Expression, Including Unexpected Hyper-Methylation at Developmental Genes

(A) Bimodal distribution of distances between hypo-DMRs in each cell type indicates that some hypo-DMRs are closely spaced (<1 kb separation) and form large blocks of differential methylation ("large hypo-DMRs").

(B) Large hypo-DMRs and a H3K4me3+ DNA methylation valley (DMV) overlap *Mef2c* (left); a H3K27me3+ DMV overlaps *Gbx2* (right). As diagrammed for the excitatory neuron tracks, dark-colored bars indicate hypo-DMRs (upper), boxes indicate hypo-DMRs that were grouped into large hypo-DMRs, and light-colored bars indicate DMVs (lower).

(C) For excitatory neurons, violin plots show the distribution of histone modification enrichments (left), ATAC-seq read densities (middle), and gene expression levels (right) within large hypo-DMRs, hypo-DMRs <2 kb, and DMVs. A.U., arbitrary units.

(D) Matrix showing the percentage of each row feature that overlaps with differentially expressed genes. Large hypo-DMRs and H3K4me3+ DMVs (in excitatory neurons) have higher enrichment for differentially expressed genes, compared to hypo-DMRs <2 kb. H3K27me3+ DMVs (in excitatory neurons) are not enriched for differentially expressed genes at q<1E-5.

(E) Schematic for assessing the accumulation of CG methylation in each adult cell type (excitatory, PV, and VIP neurons, and glia) compared to fetal cortex, at fetal DMVs overlapping genes.

(F) DNA methylation levels for a region around *Neurog2* (left), an active TF in excitatory and many glial progenitors, and *Nkx2-1* (right), a transiently active TF in PV neuron development. Barplots show the % mCG and % mCH for each cell type in the region between dotted lines in Figure 13E. * $q<1\times10^{-10}$ (mCG, adult cell type compared to fetal cortex, 1-sided FET with

Benjamini-Hochberg correction). In the browser representation, light-colored bars indicate DMVs.

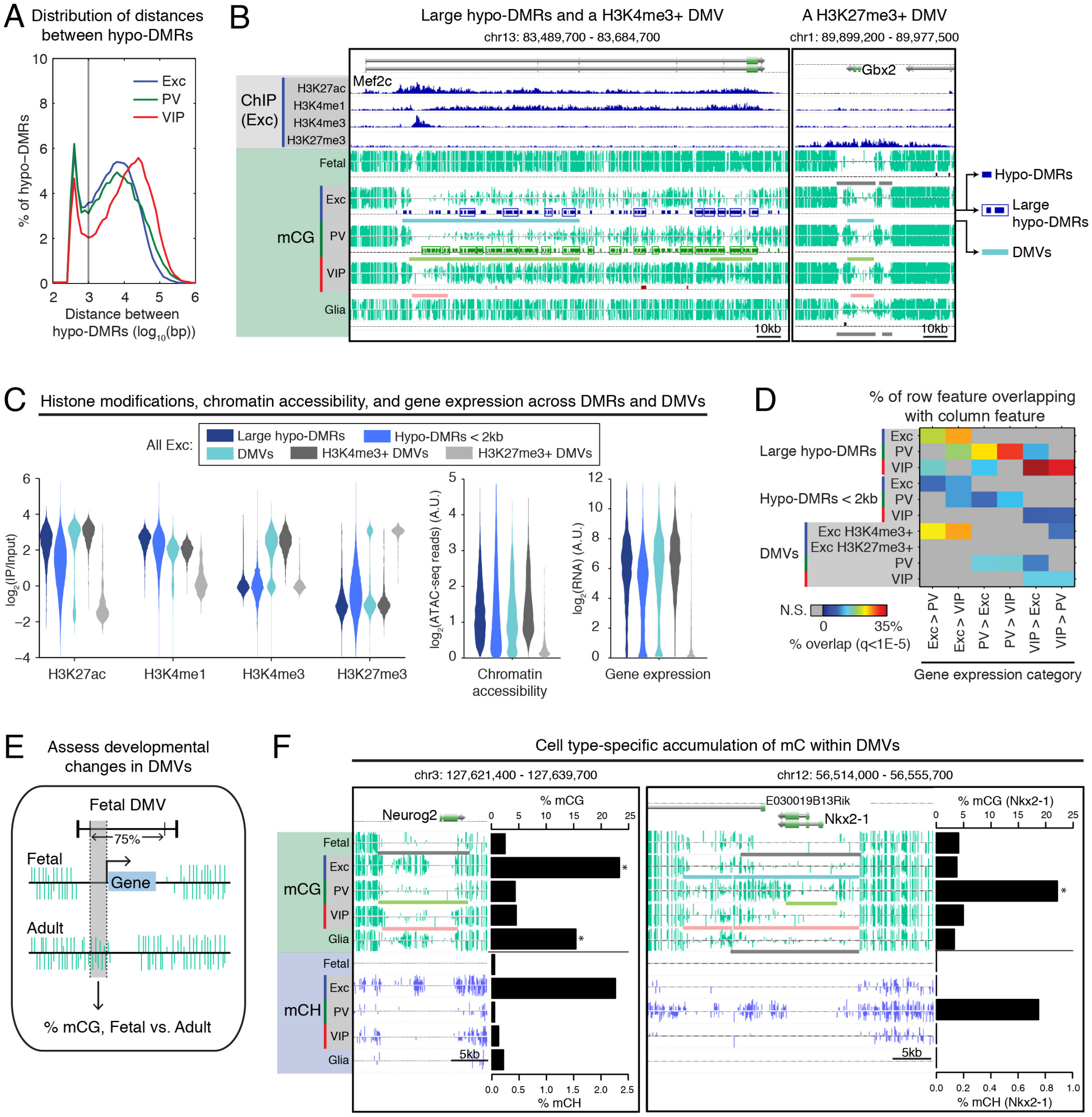


Figure 14. Large Hypo-Methylated Domains

(A) Large hypo-DMRs are generally non-overlapping across cell types, whereas DMVs show high overlap across cell types. Large hypo-DMRs and DMVs are generally non-overlapping regions in the same cell type. The numbers of large hypo-DMRs and DMVs identified in each cell type are indicated in parentheses.

(B) Boxplot showing the length distributions for large hypo-methylation features compared to all DMRs and UMRs+LMRs. Large hypo-DMRs and DMVs are both multi-kilobase DNA methylation features. By definition, the lower size limits are 2 kb for large hypo-DMRs and 5 kb for DMVs. Autosomal features for excitatory, PV, and VIP neurons were combined. Outliers are omitted in the graphical representation.

(C-D) Distribution of GC content (C) and CG methylation level (D) across DNA methylation features. Excitatory neuron features and methylation levels were used, as well as randomly selected genomic regions matching the sizes of excitatory hypo-DMRs with lengths less than 2 kb.

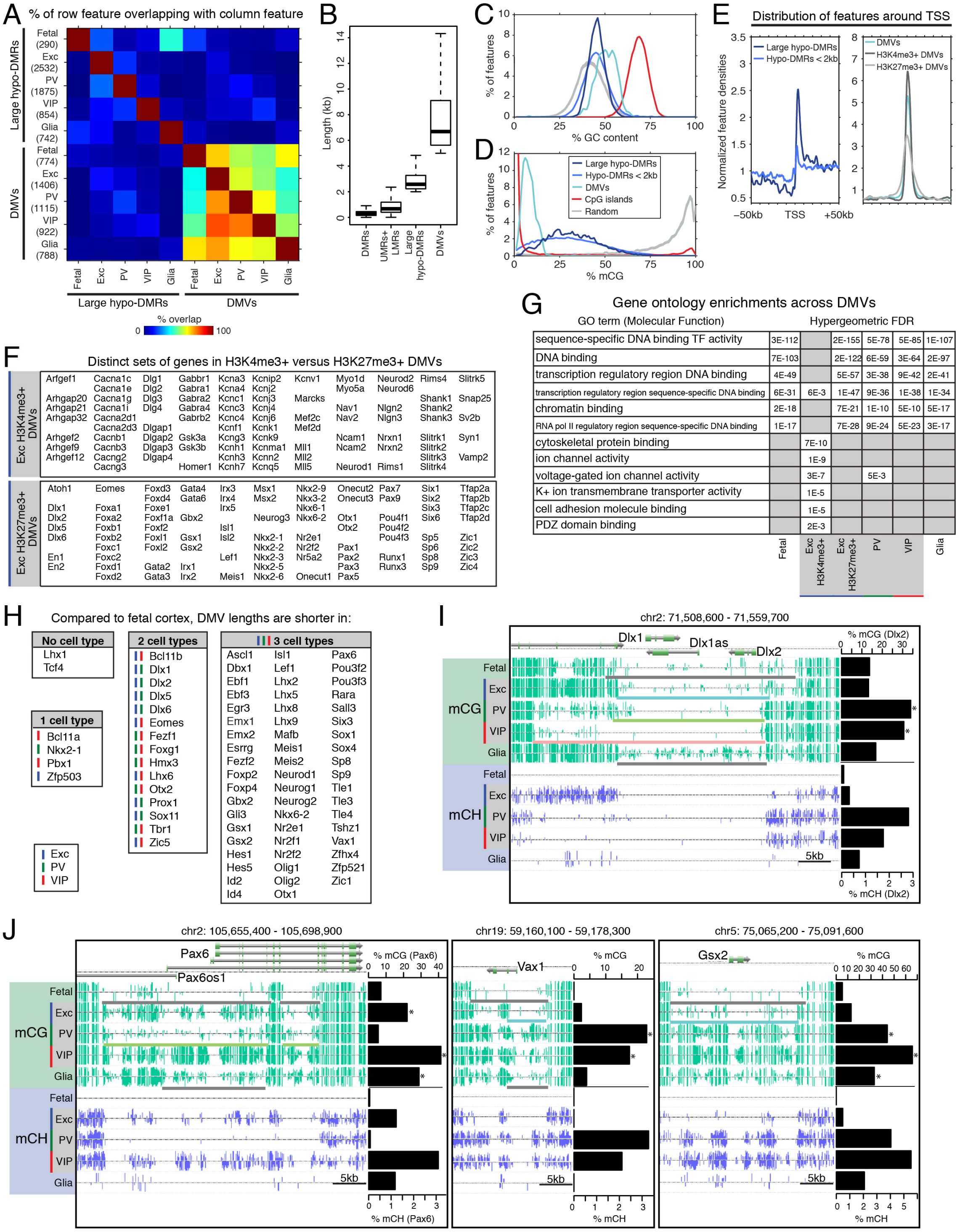
(E) Line plots showing that large hypo-DMRs are enriched downstream of the TSS whereas DMVs are enriched equally across the TSS. Excitatory neuron features were used.

(F) Representative selection of genes in excitatory DMVs that overlap H3K4me3+ peaks (top) and H3K27me3+ domains (bottom).

(G) Gene ontology (GO) categories (McLean et al., 2010) related to transcription regulation and TF activity are strongly enriched at H3K27me3+ excitatory DMVs and DMVs in other cell types, including fetal brain. H3K4me3+ excitatory DMVs are enriched for terms related to mature neuronal function.

(H) Out of 77 developmental TFs (Visel et al., 2013) that overlap fetal DMVs, the DMV lengths for 75 TFs are shorter in at least one adult cell type relative to fetal cortex. For each TF, the cell type(s) with decreased DMV length(s) are indicated.

(I-J) (I) DNA methylation levels for a region around *Dlx1/2*, showing extensive neuron subtype-specific differences in the boundaries of DMVs that correlate with developmental shifts in the expression of *Dlx2* and *Dlx1*. (J) DNA methylation levels for a region around *Pax6* (left), *Vax1* (middle), and *Gsx2* (right). *Pax6* is expressed during excitatory neuron development and in the caudal ganglionic eminence (birthplace of VIP neurons), whereas *Vax1* and *Gsx2* are expressed during inhibitory neuron development. Expression levels of all three TFs are largely down-regulated in mature neurons. For (I) and (J), barplots show the % mCG and % mCH for each cell type at the region between the dotted lines in Figure 13E. * $q<1\times10^{-10}$ (mCG, adult cell type compared to fetal cortex, 1-sided FET with Benjamini-Hochberg correction). In the browser representation, light-colored bars indicate DMVs.



F

Distinct sets of genes in H3K4me3+ versus H3K27me3+ DMVs

Exc H3K4me3+ DMVs

Arfgef1
Arhgap20
Arhgap21
Arhgap32
Arhgef2
Arhgef9
Arhgef12
Cacna1c
Cacna1e
Cacna1g
Cacna1i
Cacna2d1
Cacna2d3
Cacnb1
Cacnb3
Cacng2
Cacng3
Dlg1
Dlg2
Dlg3
Dlg4
Dlgap1
Dlgap2
Dlgap3
Dlgap4
Gabbr1
Gabra1
Gabra2
Gabra4
Gabrb2
Gsk3a
Gsk3b
Homer1
Kcna3
Kcna4
Kcnc1
Kcnc3
Kcnc4
Kcnf1
Kcng3
Kcnh1
Kcnh3
Kcnh7
Kcnip2
Kcnj2
Kcnj3
Kcnj4
Kcnj6
Kcnk1
Kcnk9
Kcnma1
Kcnn2
Kcnq5
Kcnv1
Marcks
Mef2c
Mef2d
Mll1
Mll2
Mll5
Myo1d
Myo5a
Nav1
Nav2
Ncam1
Ncam2
Neurod1
Neurod2
Neurod6
Nlgn2
Nlgn3
Nrxn1
Nrxn2
Rims1
Rims4
Shank1
Shank2
Shank3
Slitrk1
Slitrk2
Slitrk3
Slitrk4
Slitrk5
Snap25
Sv2b
Syn1
Vamp2

Exc H3K27me3+ DMVs

Atoh1
Dlx1
Dlx2
Dlx5
Dlx6
En1
En2
Eomes
Foxa1
Foxa2
Foxb1
Foxb2
Foxc1
Foxc2
Foxd1
Foxd2
Foxd3
Foxd4
Foxe1
Foxf1a
Foxf2
Foxl1
Foxl2
Gata2
Gata3
Gata4
Gata6
Gbx2
Gsx1
Gsx2
Irx1
Irx2
Irx3
Irx4
Irx5
Isl1
Isl2
Lef1
Meis1
Msx1
Msx2
Neurog3
Nkx2-1
Nkx2-2
Nkx2-3
Nkx2-5
Nkx2-6
Nkx2-9
Nkx3-2
Nkx6-1
Nkx6-2
Nr2e1
Nr2f2
Nr5a2
Onecut1
Onecut2
Onecut3
Otx1
Otx2
Pax1
Pax2
Pax3
Pax5
Pax7
Pax9
Pou4f1
Pou4f2
Pou4f3
Runx1
Runx3
Six1
Six2
Six3
Six6
Sp5
Sp6
Sp8
Sp9
Tfap2a
Tfap2b
Tfap2c
Tfap2d
Zic1
Zic2
Zic3
Zic4

G

Gene ontology enrichments across DMVs

GO term (Molecular Function)	Hypergeometric FDR Fetal	Exc H3K4me3+	Exc H3K27me3+	PV	VIP	Glia
sequence-specific DNA binding TF activity	3E-112		2E-155	5E-78	5E-85	1E-107
DNA binding	7E-103		2E-122	6E-59	3E-64	2E-97
transcription regulatory region DNA binding	4E-49		5E-57	3E-38	9E-42	2E-41
transcription regulatory region sequence-specific DNA binding	6E-31	6E-3	1E-47	9E-36	1E-38	1E-34
chromatin binding	2E-18		7E-21	1E-10	5E-10	5E-17
RNA pol II regulatory region sequence-specific DNA binding	1E-17		7E-28	9E-24	5E-23	3E-17
cytoskeletal protein binding		7E-10				
ion channel activity		1E-9				
voltage-gated ion channel activity		3E-7		5E-3		
K+ ion transmembrane transporter activity		1E-5				
cell adhesion molecule binding		1E-5				
PDZ domain binding		2E-3				

H

Compared to fetal cortex, DMV lengths are shorter in:

No cell type

Lhx1
Tcf4

1 cell type

🟥 Bcl11a
🟩 Nkx2-1
🟥 Pbx1
🟦 Zfp503

🟦 Exc
🟩 PV
🟥 VIP

2 cell types

🟦🟥 Bcl11b
🟦🟩 Dlx1
🟦🟩 Dlx2
🟦🟩 Dlx5
🟦🟩 Dlx6
🟦🟥 Eomes
🟩🟥 Fezf1
🟩🟥 Foxg1
🟩🟥 Hmx3
🟦🟥 Lhx6
🟩🟥 Otx2
🟦🟩 Prox1
🟦🟩 Sox11
🟩🟥 Tbr1
🟦🟥 Zic5

🟦🟩🟥 3 cell types

Ascl1
Dbx1
Ebf1
Ebf3
Egr3
Emx1
Emx2
Esrrg
Fezf2
Foxp2
Foxp4
Gbx2
Gli3
Gsx1
Gsx2
Hes1
Hes5
Id2
Id4
Isl1
Lef1
Lhx2
Lhx5
Lhx8
Lhx9
Mafb
Meis1
Meis2
Neurod1
Neurog1
Neurog2
Nkx6-2
Nr2e1
Nr2f1
Nr2f2
Olig1
Olig2
Otx1
Pax6
Pou3f2
Pou3f3
Rara
Sall3
Six3
Sox1
Sox4
Sp8
Sp9
Tle1
Tle3
Tle4
Tshz1
Vax1
Zfhx4
Zfp521
Zic1

Curriculum Vitae

Alisa Mo **amo4@jhmi.edu**

Personal Data

Born January 29, 1987; Chongqing, China

Education

2008 – present	Medical Scientist Training Program Johns Hopkins University School of Medicine, Baltimore, MD
2010 – 2015	Ph.D. student Neuroscience graduate program Johns Hopkins University School of Medicine, Baltimore, MD
2004 – 2008	B.A., Biological Sciences (summa cum laude) B.A., Mathematics (cum laude) Cornell University, Ithaca, NY

Research Experience

2010 – 2015	Ph.D. student with Dr. Jeremy Nathans Department of Molecular Biology and Genetics Johns Hopkins University School of Medicine, Baltimore, MD
2005 – 2008	Research assistant with Dr. Robert F. Gilmour, Jr. Division of Biomedical Sciences Cornell University, Ithaca, NY
2005 (summer)	Research assistant with Dr. Owen Obel Department of Cardiology Dallas VA Medical Center, Dallas, TX
2004 (summer)	Research assistant with Dr. Q. Richard Lu Center for Developmental Biology University of Texas Southwestern, Dallas, TX

Publications

1. **Mo*, A.**, Mukamel*, E.A., Davis*, F.P., Luo*, C., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Sejnowski, T.J., Lister, R., Eddy, S.R., Ecker, J.R., and Nathans, J.

Epigenomic signatures of neuronal diversity in the mammalian brain. (in review) (*co-first authors)

2. **Mo, A.**, Luo, C., Beer, M.A., Davis, F.P., Mukamel, E.A., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Eddy, S.R., Ecker, J.R., and Nathans, J. Epigenomic differences between retinal rods and cones reflect function and chromatin organization. (manuscript in preparation)
3. Wu, H., Luo, J., Yu, H., Rattner, A., **Mo, A.**, Wang, Y., Smallwood, P.M., Erlanger, B., Wheelan, S.J., and Nathans, J. (2014). Cellular resolution maps of X chromosome inactivation: implication for neural development, function, and disease. Neuron *81*, 103-119.
4. Rosenblatt, A., Kumar, B.V., **Mo, A.**, Welsh, C.S., Margolis, R.L., and Ross, C.A. (2012). Age, CAG repeat length, and clinical progression in Huntington's disease. Mov. Disord. *27*, 272-276.
5. Otani, N.F., **Mo, A.**, Mannava, S., Fenton, F.H., Cherry, E.M., Luther, S., and Gilmour, R.F. Jr. (2008). Characterization of multiple spiral wave dynamics as a stochastic predator-prey system. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. *78*, 021913.

Oral Presentations (as presenting author)

1. **Mo, A.**, Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Eddy, S.R., Ecker, J.R., and Nathans, J. (2014). Cell type-specific epigenetic configurations in the mammalian brain. Neuroepigenetics - Society for Neuroscience Satellite Event
2. **Mo, A.**, Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Eddy, S.R., Ecker, J.R., and Nathans, J. (2014). Unique patterns of epigenetic control in distinct subtypes of neocortical neurons. Janelia Conference on High-Throughput Sequencing for Neuroscience (selected short talk)

Grants and Fellowships

2008 – present	Medical Scientist Training Program fellowship
2011 – 2012	Visual Neuroscience Training Program fellowship

Awards and Honors

2014	Best poster presentation, Neuroscience department retreat, Johns Hopkins University School of Medicine
2014	Travel award, Graduate Student Association, Johns Hopkins University School of Medicine
2008	Merrill Presidential Scholar, Cornell University
2007	Barry M. Goldwater Scholar

Teaching Experience

2011	Teaching assistant, Neuroscience and Cognition I, Johns Hopkins University School of Medicine

Work Experience

2006 – 2008	Writing Walk-In Center Tutor, John S. Knight Institute for Writing, Cornell University

Volunteer and Leadership Activities

2015 – present	Contributor to Biomedical Odyssey, a blog sponsored by Johns Hopkins University School of Medicine
2013 – present	Volunteer for Project Bridge, Johns Hopkins University School of Medicine
2009 – 2010	Tutor at Wolfe Street Academy, Baltimore, MD
2008 – 2010	Member and Vice President, Asian Pacific American Medical Student Association, Johns Hopkins University School of Medicine
2006 – 2008	Co-founder, president, and marketing director of Cornell Health International, Cornell University
2006 – 2008	Student Advisory Board, Global Health Concentration, Cornell University