Winter 1-15-2021

Computational Analysis of Non-CpG DNA Methylation in the Mammalian Nervous System

Dennis Yawen Wu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
John Edwards, Chair
Harrison Gabel
Kristen Kroll
Tychele Turner
Ting Wang

Computational Analysis of Non-CpG DNA Methylation in the Mammalian Nervous System

by
Dennis Yawen Wu

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

January 2021
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

None of this work would have been possible without the many people supporting and advising me along the Ph.D. path. Firstly, I would like to extend my deepest gratitude to Harrison Gabel for being a heaven-sent mentor. Thank you for taking a chance on me, and letting a student with 5 rotations and many, many more failed westerns become your first graduate student. Thank you for always giving me your undivided attention, even while you quickly became the 8th floor's most popular PI. And thank you for creating a better lab than I could have dreamed of joining.

I would also like to thank each of the members of my thesis committee: John Edwards, Kristen Kroll, Ting Wang, and Tychele Turner. Thank you for investing your valuable time in my training and providing enthusiastic and critical feedback on my project. I would like to thank all of the members of the Gabel lab for being great co-workers and better friends. My thesis would not have been remotely possible without Adam Clemens and Diana Christian, who never failed to be tremendously insightful and delightful. Sabin Nettles always stood up for me and anyone else who needed it, and was the bedrock the lab needed, but didn't deserve. And thank you to Ryan, Jenna, Russell, Nicole, Chibueze; I couldn't have done this without you.

Equally important to my survival through Graduate school are all my comrades-in-arms through this journey. Jennifer Flynn, Liron Ganel, Nick Jacobs, and Rachel Lyman provided valuable insights in this project, and more importantly, were always there to make graduate school a truly wonderful experience. The real Ph.D. was the friends we made along the way.

And to my family: thank you, from the bottom of my heart, for your unyielding support and unconditional love. None of this would have been possible without you and your truly inspiring level of devotion. You've helped me every single step of the way, and you continually impress me with your dedication, wisdom, and kindness. Thank you.

Dennis Wu

Washington University

January 2021

Abstract of the Dissertation

Computational Analysis of Non-CpG DNA Methylation in the Mammalian Nervous System

by

Dennis Yawen Wu

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2020

Professor John Edwards, Chair

Noncanonical forms of DNA methylation, especially non-CpG DNA methylation, play essential roles in the neuronal epigenome, and have only recently begun to be characterized. While most DNA methylation within mammals is found in a CG context and maintained by DNMT1, neurons contain uniquely high levels of non-CpG methylation, such that the total amounts of methylation in non-CpG contexts equals or surpasses the total amounts of methylation in CG contexts. Non-CpG methylation, unlike CpG methylation, cannot be maintained by DNMT1, and must be established by the de novo methyltransferase DNMT3A.

One unique characteristic of non-CpG methylation compared to canonical CpG methylation is the extremely wide range of biological signal it exhibits across large regions of the genome. This may enable MeCP2, a critical methyl-binding protein whose disruption causes

multiple neurodevelopmental diseases, to repress regulatory elements across entire domains of the genome. How these patterns of varied methylation are established throughout the genome and what factors direct methylation to one location versus another, however, are unknown. In addition, the mechanisms by which these methyl binding proteins function and the transcriptomic effects when non-CpG methylation is lost are not well understood. As such, my dissertation work centers around the patterning, function, and consequences of this unique neural regulatory mark.

Firstly, by applying multiscale analysis of bisulfite-sequencing and high-throughput chromatin conformation capture data in the cerebral cortex of mice we find that megabase-scale regions of high non-CG methylation can correspond with topologically-associating domains of chromatin folding, identifying a new mechanism influencing mCA deposition across the neuronal genome. We find that MeCP2 represses enhancers found in these domains when they are enriched for non-CG and CG methylation, with the strongest repression occurring for enhancers located within MeCP2-repressed genes. These alterations in enhancer activity provide a mechanism for how MeCP2 disruption in neurodevelopmental disorders can lead to widespread changes in gene expression.

In light of our findings that enhancer-based repression by MeCP2 and mCA is disrupted in models of MeCP2 disorders, we investigated whether this pathway is affected in a new model of NDD caused by DNMT3A disruption.We find that multiple transcriptomic and epigenomic changes are shared between a knockout of MeCP2, and a heterozygous knockout of DNMT3A, the enzyme that establishes neuronal non-CpG methylation.

Together these findings demonstrate a previously unrecognized role for non-CpG DNA methylation in the regulation of enhancer activity in neurons, and a role for enhancer dysregulation, stemming from disruption of non-CpG DNA methylation, in multiple disorders. This highlights non-CpG methylation as a possible convergence point between multiple neurodevelopmental disorders.

Chapter 1: Introduction

The mammalian nervous system requires extremely precise control of gene expression and regulatory networks in order to develop properly, and these requirements change over each stage of development. A critical mechanism that contributes to gene regulation is DNA methylation (Bird & Wolffe, 1999; P. A. Jones & Takai, 2001), the addition of a methyl group to the 5' position of a cytosine. DNA methylation is primarily associated with gene repression (Bird & Wolffe, 1999; Greenberg & Bourc'his, 2019), as it is enriched within silenced promoters and highly repetitive regions of the genome. There are several known mechanisms by which DNA methylation can lead to gene repression, including blocking the binding of activating transcription factors (Tate & Bird, 1993) and recruiting methyl-binding proteins that can have repressive effects on transcription (P. L. Jones et al., 1998). This relationship, however, is not absolute; some transcription factors can bind to both methylated and unmethylated DNA, and methylation changes can occur downstream of activation or repression (Feldmann et al., 2013). There are developmental consequences if DNA methylation is disrupted, as zygotic ablation of DNA methylation through knockout of DNA methyltransferases results in embryonic lethality (Li et al., 1992; Okano et al., 1999), and mutations in these methyltransferases and disruption of methylation are widely associated with cancer (Das & Singal, 2004; Robertson, 2001).

DNA methylation appears to be especially important within the mammalian brain, as neuronal DNA is enriched for noncanonical forms of methylation, and disruption of DNA methyltransferase 3a (DNMT3A) within the brain leads to neurological defects in mice. Altering

methylation specifically within the mouse brain through a conditional knockout of DNMT3A leads to broad behavioral and neuromuscular defects (Nguyen et al., 2007). Within humans, mutations in a major neuronal reader of methylation, Methyl-CpG Binding Protein 2 (MeCP2) cause Rett syndrome, another severe neurological disorder (Amir et al., 1999). Recently, more links between human disease and neuronal methylation have been found, as new exome sequencing studies have identified heterozygous mutations in DNMT3A in autism spectrum disorders (Feliciano et al., 2019; Sanders et al., 2015; Satterstrom et al., 2020) and Tatton-Brown-Rahman syndrome (TBRS), a neurodevelopmental disorder typified by overgrowth and intellectual disability (Tatton-Brown et al., 2014, 2018). Given the clear links to disease and the growing amount of evidence that methylation has a particularly important role in neuronal function, there is a pressing need to understand more about the unique forms of neuronal methylation, how they are established, the functions of their readers, and the diseases that result when neuronal methylation is disrupted.

1.1 The unique neuronal methylome

While DNA methylation within mammals is typically characterized by the addition of a methyl group to a cytosine followed by a guanine (mCG), that methyl group can become oxidized and converted into a hydroxymethyl group by Tet Methylcytosine Dioxygenase 1 (TET1) (Tahiliani et al., 2009; Wu & Zhang, 2014). This form of methylation, hydroxymethylation (hmCG), is commonly thought of as an intermediary in removing methylation, through both passive and active mechanisms (He et al., 2011). While hmCG is present within most cell types, it is enriched 10-fold within neurons (Globisch et al., 2010;

Kriaucionis & Heintz, 2009) and is a stable mark that builds up within gene bodies proportionally to those genes' expression levels (Guo et al., 2014; Mellén et al., 2012). hmCG provides another possible dimension to the neuronal epigenome, as it has been suggested to functionally demethylate highly expressed genes, rendering them less repressible by methyl binding proteins such as MeCP2 (Chen et al., 2015; Gabel et al., 2015; Mellén et al., 2017).

While hmCG function is largely defined by its ability to actively, passively, and functionally demethylate regions (a sort of epigenetic "negative space"), another unique form of methylation in mammalian neurons may be able to recruit regulators in a way that traditional mCG is unable to. This unique noncanonical methylation differs by context, rather than type of modification. In most cells, the vast majority of DNA methylation is in a CG context, but neurons contain uniquely high levels of non-CpG DNA methylation (mCH, where H = A, T, or C), the majority of which is in a CA context (mCA) (Lister et al., 2013; Schultz et al., 2015). While the methylation rate (mC/C) of any given CA dinucleotide is low (the majority between 2-5% in neurons, compared to .01% or lower for most other cell types) compared to methylation at CG sites (70-90%), there are many more CA sites in mammalian genomes than CG sites (in mice, ~380 million CA sites vs ~43 million CG sites), due to the susceptibility of methylated cytosines to be deaminated and converted into a thymine over evolutionary time (Sved & Bird, 1990). Because of this, the total number of methylated sites in a CA context rivals the total number of mCG sites within adult neurons. The high quantities of mCA in neurons compared to other cell types, and its prevalence equaling that of mCG in these cells suggests that this newly discovered form of methylation may be an important epigenetic mechanism for nervous system

function (Clemens & Gabel, 2020; Lister et al., 2013). In addition, neuronal mCA can vary contiguously in regions of the genome 100kb long or larger, such that a given 100kb region of the genome can have up to three times the amount of mCA within it compared to another neighboring region. This variation may provide a wide range of biological signal for binding and repression.

Initially discovered within embryonic stem cells (Ramsahoye et al., 2000), non-CpG methylation is also enriched within heart cells and glia (Schultz et al., 2015). However, neurons contain up to ten times as much mCA as these cells, and certain neuronal subtypes contain even more mCA (Mo et al., 2015). The exact mechanisms by which neurons accumulate these exceptionally high levels of mCA are unknown, but several contributing factors have started to emerge. One such likely reason for neurons to accumulate high amounts of mCA is due to their postmitotic nature, and how mCA and mCG differ between their establishment and maintenance through cell division: mCG can be maintained by DNA methyltransferase 1 (DNMT1) through cell division, as DNMT1 binds to hemimethylated CG sites and rapidly methylates the opposing strand (Bestor, 2000; Leonhardt et al., 1992; Li et al., 1992). However, mCA is rapidly lost through cell division, as DNMT1 has no such mechanism for maintaining it. Instead, mCA must be established by the de-novo methyltransferases DNMT3A and DNMT3B (Bestor, 2000; Gabel et al., 2015; Ramsahoye et al., 2000), which, unlike DNMT1, do not depend on hemimethylation in order to efficiently deposit a mC. Thus, mCA likely has a greater ability to build up over time within neurons and other postmitotic cells, such as cardiac cells. Mammalian neurons also

experience a surge of expression of DNMT3A in early postnatal development (Feng et al., 2005), contributing to their high levels of mCA.

1.2 DNA methyltransferase 3A establishes neuronal mCA

Additional analysis of DNMT3A and mCA patterns in neurons has established that DNMT3A is the primary enzyme responsible for deposition of mCA within neurons. A neuronal-specific knockout of DNMT3A leads to the complete ablation of mCA with comparatively minor effects on mCG (Gabel et al., 2015), and analysis of DNMT3A binding at different timepoints finds that early postnatal binding of DNMT3A is highly correlated with adult mCA levels (Stroud et al., 2017). Characterizing DNMT3A and the factors that influence its binding at this time period is thus essential in determining how mCA is distributed across the neuronal genome. Few existing studies describing DNMT3A have examined its regulation in the nervous system, where DNMT3A is exceptionally highly expressed, and there are yet to be any studies that examine how the timing of DNMT3A expression and the factors that promote DNMT3A binding impact mCA.

Outside of the nervous system, peptide assays characterizing the ADD protein-interacting domain of DNMT3A have found that it binds to the N-terminal tail of Histone H3. Notably, this binding disrupted in the presence of methylated Histone H3 lysine 4 (H3K4me1/2/3)(Zhang et al., 2010), which mark active and poised transcription start sites. It is thought that this inhibition of binding contributes to a strong depletion of both mCA and mCG within these active regulatory regions. Another domain of DNMT3A, the PWWP domain, has been shown to bind to

Histone H3 lysine 36 trimethylation (H3K36me3) (Dhayalan et al., 2010), a histone mark enriched within highly transcribed gene bodies (Kolasinska-Zwierz et al., 2009). However, seemingly at odds with this finding, mCA is depleted within expressed genes in neurons, rather than enriched. More recent work in non-neural cells has suggested while DNMT3A does bind to H3K36me3, it preferentially binds to H3K36me2 with greater efficiency, while DNMT3B preferentially binds to H3K36me3 (Weinberg et al., 2019; Xu et al., 2020). These observations may have implications for mCA in the nervous system, where mCA levels are low at expressed gene bodies, where H3K36me3 is abundant, but high within broad euchromatic regions. H3K36me2 profiles are broadly euchromatic in non-neural cells, suggesting that similar patterns could be recruiting DNMT3A and driving broad mCA deposition, but to date the profiles of H3K36me2 have not been assessed in neurons. In addition, other factors such as RNA Polymerase II may also have an effect on DNMT3A binding (Stroud et al., 2017). While these factors that interact with DNMT3A may impact mCA deposition, no studies have been performed on if, how, and when they do. DNMT3A is expressed at extremely high levels in the brain (Feng et al., 2005), and its surge and subsequent decline of expression during early postnatal development suggests a potential role for mCA and DNMT3A in an important period of synaptic pruning and refinement. Exactly what this role is, and how the temporal dynamics of DNMT3A help accomplish its regulatory role, however, must be elucidated.

1.3 mCA forms the basis of a critical neural-specific epigenetic mechanism

DNMT3A mutations in humans lead to Tatton-Brown-Rahman syndrome (Tatton-Brown et al. 2014), and have been linked to autism spectrum disorders. Recent work characterizing the

function of DNMT3A and mCA through epigenomic and transcriptomic studies reveals a possible link to Rett syndrome and other neurodevelopmental diseases. Studying the gene expression changes of the aforementioned neuronal-specific knockout of DNMT3A identifies hundreds if not thousands of subtly changed genes (Gabel et al., 2015). The genes most changed upon loss of Dnmt3a have high amounts of mCA throughout their gene bodies (Gabel et al., 2015), a characteristic shared with genes changed in a mouse model of Rett syndrome, suggesting potentially overlapping mechanisms and pathology though a prominent neuronal reader of methylation, MeCP2.

MeCP2, similarly to DNMT3A and mCA, is highly enriched within neurons, where its expression reaches near-histone levels (Skene et al., 2010). While it was first identified through its strong affinity to mCG, further research has shown it also binds tightly to mCA (Buchmuller et al., 2020; Gabel et al., 2015; Guo et al., 2014; Lagger et al., 2017). Disruptions of MeCP2 in humans cause Rett syndrome, a severe neurological disorder that is characterized by a period of normal development from birth to 6-18 months, then a gradual loss of speech followed by microcephaly, ataxia, seizures, and autism-like symptoms (Amir et al., 1999). This onset of symptoms takes place concurrently with the rise of expression in MeCP2 as well as mCA deposition, suggesting that the loss of MeCP2 interferes with proper postnatal neurological function. However, it has been difficult to determine how MeCP2 alters transcription, and what its targets are. Along with its histone-like expression levels within neurons, it exhibits extremely broad, histone-like binding patterns, with slight enrichments within highly methylated regions, but no sharp peaks to identify discrete binding sites with (Buchmuller et al., 2020; Lagger et al.,

2017; Skene et al., 2010). In MeCP2 mutants in mice, hundreds or thousands of genes experience subtle (<2-fold) changes in expression, both up and down (Chahrour et al., 2008; Gabel et al., 2015).

Although few trends have been found that explain the genes downregulated in the knockout (MeCP2-activated genes), several striking trends have been detected in the genes upregulated in the MeCP2 knockout (MeCP2-repressed genes) that link the transcriptomic changes to the unique neuronal methylome. Firstly, although these MeCP2-repressed genes do not display an enrichment of mCA or mCG at their promoters, the canonical site of action for repression and activation of gene expression, they are enriched for mCA within their gene bodies. Secondly, MeCP2-repressed genes have long transcribed gene lengths, many of them spanning over 100kb in size (Gabel et al., 2015; Kinde et al., 2016). Thus, long, highly methylated genes that are particularly susceptible to repression by MeCP2 may require the specific patterns of neuronal mCA, which can vary greatly over large genomic regions, unlike mCG. In addition, recent work studying mice carrying an engineered MeCP2 protein that can bind mCG but not mCA has demonstrated the critical importance of mCA to MeCP2 function. These mice recapitulate many of the same phenotypes seen in a complete MeCP2 KO, indicating that MeCP2 requires mCA in order to properly guide neuronal function (Tillotson et al., 2020). While these findings demonstrate the importance of mCA to regulation by MeCP2, many questions remain unanswered. How do large regions of the genome maintain and establish varied methylation levels, such that genes over 100kb long can have consistent levels of high mCA? Do the high mCA levels within these genes impact MeCP2 binding and regulation? And if MeCP2

does rely on high mCA levels at the gene body to bind to, how does MeCP2 then affect the expression levels of that gene?

1.4 Disruptions in multiple mCA-related proteins result in disease

The critical importance of mCA to neuronal function is illustrated clearly from the overlapping neurological deficits resulting from disrupting its deposition through DNMT3A or from disrupting its readout by MeCP2. However, it is unknown what role methylation changes play in TBRS, and how mCA and MeCP2 interact to regulate genes. The recent discovery of links between neuronal mCA and disease emphasizes the need to learn more about the origins and factors influencing this novel form of methylation. Mutations in Nuclear receptor SET Domain-containing protein 1 (NSD1), which establishes H3K26me2, are known to cause Sotos syndrome, a disease with features of overgrowth and intellectual disability similar to TBRS (Kurotaki et al., 2002). Given that H3K36me2 has recently been proposed to recruit DNMT3A outside of neurons, it is possible that the overlapping features between Sotos syndrome and TBRS are partially due to changes in neuronal mCA. Lending further support for this hypothesis is the finding that patients with Sotos syndrome experience numerous methylation changes at CG sites within blood (Choufani et al., 2015), showing that NSD1 mutations can cause methylation changes, and may potentially affect mCA within the brain. The commonalities between Sotos syndrome, TBRS, and Rett suggest that the unique neuronal factor that these three disorders share, mCA, is critical for normal neuronal function, and that disruptions at multiple steps in this neuronal-specific regulatory mechanism can result in disease. Exactly how mCA regulates

neuronal expression, and how disruptions in mCA depositors and readers impact the nervous system, however, is yet to be known.

To provide insight on these questions, I present my research on insights into how genomic topology influences mCA patterns to remain at a consistent level across long stretches of the genome. I subsequently show that these insights shed light on the mechanism by which MeCP2 represses long genes with high methylation, revealing that this process occurs through repression of highly-methylated intragenic enhancers. In my third chapter, I will present work on the transcriptomic and epigenetic changes in a new mouse model of TBRS. Therein, I will present evidence that there is a global loss of mCA upon heterozygous disruption of DNMT3A, and that transcriptomic changes overlap between mouse models of Rett syndrome and TBRS.

1.5 References Cited

1. Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, *23*(2), 185–188.
2. Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Human Molecular Genetics*, *9*(16), 2395–2402.
3. Bird, A. P., & Wolffe, A. P. (1999). Methylation-induced repression—belts, braces, and chromatin. *Cell*, *99*(5), 451–454.
4. Buchmuller, B. C., Kosel, B., & Summerer, D. (2020). Complete Profiling of Methyl-CpG-Binding Domains for Combinations of Cytosine Modifications at CpG Dinucleotides Reveals Differential Read-out in Normal and Rett-Associated States. *Scientific Reports*, *10*(1), 4053.
5. Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T. C., Qin, J., & Zoghbi, H. Y. (2008). MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science*, *320*(5880), 1224–1229.
6. Chen, L., Chen, K., Lavery, L. A., Baker, S. A., Shaw, C. A., Li, W., & Zoghbi, H. Y. (2015). MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5509–5514.
7. Choufani, S., Cytrynbaum, C., Chung, B. H. Y., Turinsky, A. L., Grafodatskaya, D., Chen, Y. A., Cohen, A. S. A., Dupuis, L., Butcher, D. T., Siu, M. T., Luk, H. M., Lo, I. F. M., Lam, S. T. S., Caluseriu, O., Stavropoulos, D. J., Reardon, W., Mendoza-Londono, R., Brudno, M., Gibson, W. T., … Weksberg, R. (2015). NSD1 mutations generate a genome-wide DNA methylation signature. *Nature Communications*, *6*, 10207.
8. Clemens, A. W., & Gabel, H. W. (2020). Emerging Insights into the Distinctive Neuronal Methylome. *Trends in Genetics: TIG*. https://doi.org/10.1016/j.tig.2020.07.009
9. Clemens, A. W., Wu, D. Y., Moore, J. R., Christian, D. L., Zhao, G., & Gabel, H. W. (2020). MeCP2 Represses Enhancers through Chromosome Topology-Associated DNA Methylation. *Molecular Cell*, *77*(2), 279–293.e8.
10. Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *22*(22), 4632–4642.
11. Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R. Z., Ragozin, S., & Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of Biological Chemistry*, *285*(34),

26114–26120.
12. Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., & Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genetics*, *9*(12), e1003994.
13. Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T. N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L. G., Muzny, D. M., Sabo, A., SPARK Consortium, Gibbs, R. A., Eichler, E. E., O'Roak, B. J., Michaelson, J. J., Volfovsky, N., Shen, Y., & Chung, W. K. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genomic Medicine*, *4*, 19.
14. Feng, J., Chang, H., Li, E., & Fan, G. (2005). Dynamic expression of de novo DNA methyltransferases Dnmt3a and Dnmt3b in the central nervous system. *Journal of Neuroscience Research*, *79*(6), 734–746.
15. Gabel, H. W., Kinde, B., Stroud, H., Gilbert, C. S., Harmin, D. A., Kastan, N. R., Hemberg, M., Ebert, D. H., & Greenberg, M. E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, *522*(7554), 89–93.
16. Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., & Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS One*, *5*(12), e15367.
17. Greenberg, M. V. C., & Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews. Molecular Cell Biology*, *20*(10), 590–607.
18. Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G.-L., & Song, H. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature Neuroscience*, *17*(2), 215–222.
19. He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C., & Xu, G.-L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, *333*(6047), 1303–1307.
20. Jones, P. A., & Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, *293*(5532), 1068–1070.
21. Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., & Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, *19*(2), 187–191.
22. Kinde, B., Wu, D. Y., Greenberg, M. E., & Gabel, H. W. (2016). DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(52), 15114–15119.

23. Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., & Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, *41*(3), 376–381.
24. Kriaucionis, S., & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, *324*(5929), 929–930.
25. Kurotaki, N., Imaizumi, K., Harada, N., Masuno, M., Kondoh, T., Nagai, T., Ohashi, H., Naritomi, K., Tsukahara, M., Makita, Y., Sugimoto, T., Sonoda, T., Hasegawa, T., Chinen, Y., Tomita Ha, H.-A., Kinoshita, A., Mizuguchi, T., Yoshiura Ki, K.-I., Ohta, T., … Matsumoto, N. (2002). Haploinsufficiency of NSD1 causes Sotos syndrome. *Nature Genetics*, *30*(4), 365–366.
26. Lagger, S., Connelly, J. C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B. H., Yu, M., He, C., Sanguinetti, G., Sowers, L. C., Walkinshaw, M. D., & Bird, A. (2017). MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genetics*, *13*(5), e1006793.
27. Leonhardt, H., Page, A. W., Weier, H. U., & Bestor, T. H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, *71*(5), 865–873.
28. Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*(6), 915–926.
29. 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., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., … Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, *341*(6146), 1237905.
30. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S., & Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, *151*(7), 1417–1430.
31. Mellén, M., Ayata, P., & Heintz, N. (2017). 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(37), E7812–E7821.
32. 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., & Nathans, J. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, *86*(6), 1369–1384.
33. Nguyen, S., Meletis, K., Fu, D., Jhaveri, S., & Jaenisch, R. (2007). Ablation of de novo DNA methyltransferase Dnmt3a in the nervous system leads to neuromuscular defects and shortened lifespan. *Developmental Dynamics: An Official Publication of the*

American Association of Anatomists, *236*(6), 1663–1676.
34. Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257.
35. Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., & Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5237–5242.
36. Robertson, K. D. (2001). DNA methylation, methyltransferases, and cancer. *Oncogene*, *20*(24), 3139–3155.
37. Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., Goldberg, A. P., Jinlu, C., Keaney, J. F., 3rd, Klei, L., Mandell, J. D., Moreno-De-Luca, D., Poultney, C. S., Robinson, E. B., Smith, L., … State, M. W. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, *87*(6), 1215–1233.
38. Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568–584.e23.
39. Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W., & Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, *523*(7559), 212–216.
40. Skene, P. J., Illingworth, R. S., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., Andrews, R., & Bird, A. P. (2010). Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Molecular Cell*, *37*(4), 457–468.
41. Stroud, H., Su, S. C., Hrvatin, S., Greben, A. W., Renthal, W., Boxer, L. D., Nagy, M. A., Hochbaum, D. R., Kinde, B., Gabel, H. W., & Greenberg, M. E. (2017). Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell*, *171*(5), 1151–1164.e16.
42. Sved, J., & Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(12), 4692–4696.
43. Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., & Rao, A. (2009). Conversion of

5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, *324*(5929), 930–935.
44. Tate, P. H., & Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development*, *3*(2), 226–231.
45. Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., Del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., Baralle, D., Dabir, T., Gener, B., Goudie, D., Homfray, T., Kumar, A., Pilz, D. T., Selicorni, A., Temple, I. K., … Rahman, N. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. *Nature Genetics*, *46*(4), 385–388.
46. Tatton-Brown, K., Zachariou, A., Loveday, C., Renwick, A., Mahamdallie, S., Aksglaede, L., Baralle, D., Barge-Schaapveld, D., Blyth, M., Bouma, M., Breckpot, J., Crabb, B., Dabir, T., Cormier-Daire, V., Fauth, C., Fisher, R., Gener, B., Goudie, D., Homfray, T., … Rahman, N. (2018). The Tatton-Brown-Rahman Syndrome: A clinical study of 55 individuals with de novo constitutive DNMT3A variants. *Wellcome Open Research*, *3*, 46.
47. Tillotson, R., Cholewa-Waclaw, J., Chhatbar, K., Connelly, J., Kirschner, S. A., Webb, S., Koerner, M. V., Selfridge, J., Kelly, D., De Sousa, D., Brown, K., Lyst, M. J., Kriaucionis, S., & Bird, A. (2020). Neuronal non-CG methylation is an essential target for MeCP2 function. In *bioRxiv* (p. 2020.07.02.184614). https://doi.org/10.1101/2020.07.02.184614
48. Weinberg, D. N., Papillon-Cavanagh, S., Chen, H., Yue, Y., Chen, X., Rajagopalan, K. N., Horth, C., McGuire, J. T., Xu, X., Nikbakht, H., Lemiesz, A. E., Marchione, D. M., Marunde, M. R., Meiners, M. J., Cheek, M. A., Keogh, M.-C., Bareke, E., Djedid, A., Harutyunyan, A. S., … Lu, C. (2019). The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature*, *573*(7773), 281–286.
49. Wu, H., & Zhang, Y. (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, *156*(1-2), 45–68.
50. Xu, W., Li, J., Rong, B., Zhao, B., Wang, M., Dai, R., & Chen, Q. (2020). DNMT3A reads and connects histone H3K36me2 to DNA methylation. *Protein & Cell*. https://link.springer.com/article/10.1007/s13238-019-00672-y
51. Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., & Jeltsch, A. (2010). Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Research*, *38*(13), 4246–4253.

Chapter 2: MeCP2 represses enhancers through chromosome topology-associated DNA methylation

This chapter is adapted from a manuscript published in Molecular Cell:

Clemens AW, Wu DY, Moore JR, Christian DL, Zhao G, Gabel HW. MeCP2 Represses Enhancers through Chromosome Topology-Associated DNA Methylation. Mol Cell. 2020;77(2):279-293.e8. doi:10.1016/j.molcel.2019.10.033

A.W.C. carried out all of the experiments in collaboration with J.R.M. and D.L.C. and performed ChIP-seq analysis; D.Y.W. performed RNA-seq, Bisulfite-seq, and TAD analysis, and contributed to analysis of ChIP-seq studies; G.Z. and D.Y.W. performed the Hi-C analysis; A.W.C, D.Y.W, and H.W.G. designed the experiments and analysis, and wrote the manuscript.

2.1 Abstract

The genomes of mammalian neurons contain uniquely high levels of non-CG DNA methylation that can be bound by the Rett syndrome protein, MeCP2, to regulate gene expression. How patterns of non-CG methylation are established in neurons and the mechanism by which this methylation works with MeCP2 to control gene expression is unclear. Here we find that genes repressed by MeCP2 are often located within megabase-scale regions of high

non-CG methylation that correspond with topologically-associating domains of chromatin folding. MeCP2 represses enhancers found in these domains that are enriched for non-CG and CG methylation, with the strongest repression occurring for enhancers located within MeCP2-repressed genes. These alterations in enhancer activity provide a mechanism for how MeCP2 disruption in disease can lead to widespread changes in gene expression. Hence, we find that DNA topology can shape non-CG DNA methylation across the genome to dictate MeCP2-mediated enhancer regulation in the brain.

2.2 Introduction

The development and function of the mammalian brain requires precise control of gene expression (Cholewa-Waclaw et al., 2016). While DNA methylation at CG dinucleotides is used to control genes in all cells, neurons also utilize a unique form of non-CG DNA methylation for gene regulation (Lister et al., 2013; Xie et al., 2012). Non-CG DNA methylation is deposited by DNA methyltransferase 3A (DNMT3A), accumulating specifically in neurons postnatally until methylated cytosine (mC) at non-CG sites is nearly equal to mC at CG sites (mCG) (Lister et al., 2013). DNMT3A disruption in the mouse brain results in ablation of mC at non-CG sites and neurological phenotypes (Nguyen et al., 2007), supporting the importance of this methyl-mark.

Studies indicate that Methyl-CpG binding Protein 2 (MeCP2) has high affinity for mC in CA dinucleotides (mCA) similar to its classical substrate, mCG (Chen et al., 2015; Gabel et al., 2015; Guo et al., 2014). MeCP2 is expressed at near-histone levels in neurons and binds widely

across the genome, with some enrichment at regions high in mCG and mCA sites (Chen et al., 2015; Gabel et al., 2015; Skene et al., 2010). The protein interacts with the NCoR-HDAC3 corepressor complex and can block reporter expression *in vitro*, suggesting an important function of MeCP2 is gene repression (Lyst and Bird, 2015). Loss of MeCP2 causes the neurological disorder Rett syndrome, and duplication of MeCP2 leads to autism spectrum disorder (Amir et al., 1999; Van Esch et al., 2005). Transcriptomic studies of brains from Rett syndrome patients and MeCP2 knockout (MeCP2 KO) or overexpression (MeCP2 OE) mice detect only subtle changes across many genes however, making it difficult to differentiate direct targets of MeCP2 regulation from secondary gene expression effects (Chahrour et al., 2008; Tudor et al., 2002). Defining how MeCP2 mediates its broad, subtle expression effects is thus recognized as an important challenge to address (Ip et al., 2018; Lyst and Bird, 2015).

Recent analysis has revealed that MeCP2 represses genes marked by high mCA levels (Chen et al., 2015; Gabel et al., 2015; Lagger et al., 2017). Notably, these genes are not substantially enriched for mCA at promoters, a canonical site of action for methylation, but show high mCA in their transcribed regions ("gene bodies") and flanking sequences. In addition, MeCP2 KO causes a relative upregulation of genes genome-wide that is correlated with both the length and mC level of the gene body (Gabel et al., 2015; Sugino et al., 2014). These observations have led to the proposal that MeCP2 regulates expression by binding to mC in gene bodies to repress transcription (Kinde et al., 2016).

While these findings provide insight into mCA and MeCP2 in the brain, it is not known how high mCA levels are established at MeCP2-repressed genes. In addition, while MeCP2 represses the mRNA of long, high-mCA genes, it has not been demonstrated that this occurs

through direct repression of transcription, and alternative post-transcriptional mechanisms have been proposed (Johnson et al., 2017). Finally, the mechanism by which binding of MeCP2 to mC outside of promoters can drive subtle, but critical, gene repression is unknown.

Here we examined how high mCA levels are established at MeCP2-repressed genes and explored how MeCP2 functions with mCA to control transcription. We find that mCA patterning in neurons is associated with chromatin folding, and that genes most strongly repressed by MeCP2 often land in mCA-enriched topologically-associating domains (TADs). We uncover evidence that MeCP2 binds to mCA and mCG within mCA-enriched TADs to repress activity of enhancer elements and cause downregulation of promoter activity for target genes. These findings provide insight into how disruption of MeCP2 drives neurological dysfunction in Rett syndrome and related disorders.

2.3 Results

To explore how high levels of mCA are established at MeCP2-repressed genes and to determine how MeCP2 reads out mC to control gene expression, we analyzed the cerebral cortex, a brain region where MeCP2 mutations disrupt gene expression and physiology (Pacheco et al., 2017; Shepherd and Katz, 2011). We integrated our data with epigenomic datasets for this brain region (Dixon et al., 2012; Sloan et al., 2016; Stroud et al., 2017) to search for features that can explain patterns of high mCA at genes repressed by MeCP2. We focused on mCA, the most prevalent, highest affinity non-CG site for MeCP2 (Gabel et al., 2015), but obtained similar results for minor non-CG methylation sites (mCT, mCC) in our studies.

Combined RNA-seq analysis of cortex from MeCP2 KO and MeCP2 OE mice (see methods) identified 884 "MeCP2-repressed genes" significantly upregulated in MeCP2 KO and downregulated in MeCP2 OE and 843 "MeCP2-activated genes" affected in the opposite manner in these mutants (Figure 2.1A). These genes overlap with genes identified by meta-analysis of non-cortical brain regions in MeCP2 mutants (Gabel et al., 2015), indicating many MeCP2-regulated genes are shared across brain regions (Figure 2.8A,B). As in other studies (Gabel et al., 2015; Renthal et al., 2018; Sugino et al., 2014), cortical MeCP2-repressed genes are longer than the genome average (Figure 2.8D) and are enriched for mCA (mCA/CA, or "mCA level") in and around the gene, with little mCA enrichment at their transcription start site (TSS) (Figure 2.1B,C; 2.8C). We also detected global length- and mCA-associated upregulation of genes in MeCP2 KO and downregulation in MeCP2 OE (Figure 2.8E,F), with dysregulation most correlated with gene body mCA levels compared to gene-flanking regions (Figure 2.8G,H). These results confirm that MeCP2 represses long, high-mCA genes and provide a high-confidence set of MeCP2-regulated cortical genes for our analyses.

Large-scale non-CG DNA methylation profiles are linked to chromatin topology.

We next explored how high mCA levels are established at MeCP2-repressed genes. We observed that mCA enrichment can extend for megabases around these genes (Figure 2.1B,C) (Kinde et al., 2016), suggesting that regional mCA varies on a megabase scale. High-throughput chromatin conformation capture (Hi-C) analysis of the mouse cortex (Dixon et al., 2012, 2016) has detected TADs on a similar megabase scale. TADs are regions of cis-interactions, where enhancers interact with promoters in the same TAD more often than with promoters outside of

the TAD (Spielmann et al., 2018). We noted instances where high mCA surrounding MeCP2-repressed genes dropped off at transition points in Hi-C interactions similar to those delineating TADs (Figure 2.1D). A correlation between mCA patterns and TADs could have important implications for regulation of gene expression within each TAD. We therefore investigated if TADs in the cerebral cortex (Dixon et al., 2012; Weinreb and Raphael, 2016, see methods) delineate regions of high mCA at MeCP2-repressed genes.

We find that TADs containing MeCP2-repressed genes are significantly enriched for mCA compared to TADs genome-wide (Figure 2.1E). Further, we observe a drop-off of mCA levels at boundaries of TADs containing MeCP2-repressed genes that is not detected in control TADs generated by shuffling genomic locations (Figure 2.1F; 2.9B,C). While the shift in mCA across all boundaries returns toward the genome average (Figure 2.1F), a dramatic drop-off in mCA at boundaries is detected when high-mCA TADs containing MeCP2-repressed genes are adjacent to lower mCA regions (Figure 2.1G). Such strong mCA enrichments and drop-offs at boundaries are not observed for shuffled TADs (Figure 2.1F,G; 2.9B, see methods) or for potential sources of sequencing bias (Figure 2.9B), indicating that these mCA patterns are not due to technical artifacts. Cross-correlation analysis (Rao et al., 2014, see methods) shows that mCA levels are more highly correlated for regions in the same TAD than for regions in neighboring TADs (Figure 2.1H). This relationship is detected for all TADs but is most prominent for TADs containing MeCP2-repressed genes. Control TADs shuffled across the genome do not show similar correlations (Figure 2.1H, see methods). These findings suggest that TADs are units of organization for mCA levels in the genome and that MeCP2-repressed genes often occupy high-mCA TADs.

In contrast to MeCP2-repressed genes, TADs containing MeCP2-activated genes display lower mCA levels (Figure 2.1E; 2.9A). Genomic mCG levels do not show prominent megabase-scale variation, but some mCG depletion occurs in TADs containing MeCP2-activated genes (Figure 2.1D; 2.9A,B). Increased correlation of mCG within TADs is observed (Figure 2.9D), suggesting TADs do organize modest regional mCG fluctuations.

As an added test of the association between topology and regional mCA, we analyzed Hi-C data from the embryonic cortex (Bonev et al., 2017), prior to the deposition of mCA. Topology at this time point is similarly associated with adult mCA (Figure 2.9E-H), indicating that chromatin folding early in development is predictive of adult mCA patterns. We further generalized our findings using an independent cortical mC dataset (Lister et al., 2013), as well as Hi-C and mC data from the cerebellum (Mellén et al., 2017; Yamada et al., 2019). This analysis indicated that the association between genome topology and mCA levels is robustly detected in independent datasets and brain regions (Figure 2.9I-K).

We next explored how mCA is established in TADs during development. DNMT3A deposits mCA in mouse cortex from birth to six weeks of age (Lister et al., 2013), and DNMT3A binding in kilobase-scale regions at two weeks predicts mCA levels at eight weeks (Figure 2.10A; Stroud et al., 2017). We therefore asked if megabase-scale DNMT3A binding at two weeks supports a role for TADs in influencing large-scale mCA patterning. We find that DNMT3A ChIP-seq signal at two-weeks shows higher cross-correlations within TADs than across TAD boundaries (Figure 2.2A) and predicts TAD mCA in the adult cortex (Figure 2.2B). DNMT3A ChIP also drops off at boundaries of TADs containing MeCP2-repressed genes (Figure 2.2C), while input controls do not (Figure 2.10B). Analysis of topology from embryonic

neurons and DNMT3A at a later timepoint yielded similar results (Figure 2.10C-F, see methods). These findings indicate that DNMT3A binding during postnatal development is shaped by TAD structures, defining a long-term mCA "set-point" for each TAD into adulthood.

To examine how TAD mCA affects MeCP2 binding, we quantified MeCP2 ChIP-seq signal (Kinde et al., 2016) in the adult cortex. MeCP2 ChIP signal is extremely broad, showing only very modest fluctuations across the genome (Chen et al., 2015; Gabel et al., 2015). In the context of these small effects however, we find that MeCP2 levels in TADs associate with TAD mCA set-points (Figure 2.2D) and that MeCP2 cross-correlation is higher within TADs than across TAD boundaries (Figure 2.10G). Analysis of MeCP2 binding when mCA is blocked by deletion of DNMT3A (Figure 2.10H) (Kinde et al., 2016) revealed relative reductions in MeCP2 TAD binding that correlate with the wild-type mCA levels of each TAD (Figure 2.10I). Thus, high TAD mCA levels are associated with some enrichment of MeCP2 binding, and when mCA is lost, large-scale MeCP2 profiles are measurably altered.

We then considered how TAD-associated mCA levels could impact gene regulation. Methylation levels for regulatory elements and genes are thought to be determined by local sequence features and activity level at these sites (Schübeler, 2015). However, the TAD mCA set-point could act on top of these local determinants to drive consistently higher or lower mCA across all elements within each TAD. Indeed, we detect robust correlations between the TAD set-point mCA level and mCA levels of gene bodies and enhancers inside each TAD (Figure 2.2E). This correlation occurs only for elements found within the TAD, breaking down immediately outside of the TAD (Figure 2.10J). Correlations in mCA levels between elements in the same TAD are also stronger than for elements on different sides of a TAD boundary (Figure

2.10K). In contrast, mCA levels in TSS regions have limited correlation with TAD mCA levels, suggesting that these elements escape the TAD mCA set-point. Control analysis randomizing TAD or genomic element locations (see methods) confirmed that TAD boundaries delineate transitions in mCA set-points for enhancers and genes (Figure 2.10J,K) and showed that only true TSS regions escape TAD mCA levels (Figure 2.10L,M). mCG levels at these elements did not show strong TAD-associated signals (Figure 2.10L,M). These findings indicate that mCA levels at enhancers and genes in each TAD are linked to the mCA set-point for the TAD. By influencing mCA at genes and enhancers, TAD-associated mCA levels can directly impact regulation of genes within each TAD by MeCP2.

Loss of MeCP2 leads to promoter activation at MeCP2-repressed genes.

We next investigated how MeCP2 reads out mCA in TADs to affect transcription. Consistent with analysis above (Figure 2.2E), the high TAD mCA set-point for MeCP2-repressed genes is associated with high mCA in and around these genes but little mCA enrichment at the TSS (Figure 2.1B; 2.8C). This suggests that MeCP2 binds to mCA outside of the promoter to repress these genes. Based on the long length of MeCP2-repressed genes and the correlation between repression and gene body mCA (Figure 2.8D,G,H), we considered several possible regulatory mechanisms for MeCP2. For example, MeCP2 binding to gene body mCA might block transcribing RNA polymerase, resulting in premature termination and reduced mRNA. Alternatively, binding of MeCP2 to mCA at regulatory elements could act at a distance to block promoter activation. To test these and other possible mechanisms, we carried out

genomic analysis of RNA intermediates and histone marks associated with transcription in MeCP2 mutant mice.

We first performed RNA-seq on nuclear RNA from MeCP2 KO and wild-type cortex to enrich for pre-mRNA (Figure 2.3A) and analyzed intronic reads to assess changes in transcription (Boswell et al., 2017). We reasoned that if MeCP2 normally reduces the processivity of RNA polymerase II to cause premature termination of repressed genes, increased transcript completion in the MeCP2 KO would result in more intronic reads at the 3' end of genes but no change in reads at the 5' end. Alternatively, if MeCP2 represses promoter activation, loss of MeCP2 would cause a consistent increase in intronic reads along the entire gene. Finally, if post-transcriptional mechanisms affect these genes, no intronic changes would be seen. Differential expression analysis of introns detected highly concordant effects with mRNA from whole cells, with introns also showing upregulation of long and high-mCA genes relative to shorter and lower-mCA genes (Figure 2.3B; Figure 2.11A-F). Analysis of intronic RNA changes across the length of significantly upregulated genes detected a consistent increase of intronic reads across the pre-mRNA (Figure 2.3C; 2.11G,H). These results suggest that subtle changes in mRNA for MeCP2-repressed genes result from concomitant subtle changes in pre-mRNA transcription and that loss of MeCP2 leads to promoter activation, rather than changes in polymerase processivity.

As an independent measure of transcriptional effects, we performed ChIP-seq of histone modifications that report on transcriptional activity. Analysis of gene body Histone 3 lysine 36 trimethylation (H3K36me3), a mark associated with transcription levels (Guenther et al., 2007), revealed a subtle but significant increase in signal for MeCP2-repressed genes in the MeCP2 KO

(Figure 2.3A,D). Analysis of Histone 3 lysine 27 acetylation (H3K27ac) and Histone 3 lysine 4 trimethylation (H3K4me3), marks associated with promoter activation (Heintzman et al., 2007; Santos-Rosa et al., 2002), showed promoter upregulation for MeCP2-repressed genes (Figure 2.3A,D). While subtle, these effects are consistent with the magnitude of mRNA increases for these genes. Together with our intronic RNA analysis, these findings support a model in which MeCP2 acts at a distance to repress promoter activity.

Given the small magnitude of these effects, we sought to independently test if MeCP2 controls promoter activation. Based on the opposite effects on mRNA in the MeCP2 OE compared to the MeCP2 KO, we predicted intronic RNA and histone marks should be reciprocally affected in the MeCP2 OE. Indeed, integrated RNA-seq and ChIP-seq analyses in the MeCP2 OE revealed opposite effects of those in the MeCP2 KO (Figure 2.3E-G; 2.11I). These findings provide further support that MeCP2 represses long, high-mCA genes through promoter downregulation.

MeCP2 represses enhancers that are enriched for mCA and mCG binding sites.

We next considered how binding of MeCP2 to high levels of mC found outside of the TSS could lead to promoter downregulation. One way MeCP2 could achieve this is through enhancers. Since our analysis shows that high-mCA set-points of TADs containing MeCP2-repressed genes would lead to high mCA at enhancers in the TAD (Figure 2.1E; 2.2E), we considered that repression of these enhancers by MeCP2 could underlie regulation of MeCP2-repressed genes. To date however, effects of MeCP2 on enhancer activity have not been examined. Enhancer activation is associated with H3K27ac signal at these sites (Creyghton et al.,

2010), and visualization of our H3K27ac ChIP signal at enhancers near MeCP2-repressed genes suggests subtly increased acetylation in the MeCP2 KO (Figure 2.4A). Combined differential analysis of H3K27ac signal in MeCP2 KO and OE cortex genome-wide (see methods) identified significantly altered enhancers (Figure 2.4B; 2.12A). Quantification of mCA and mCG at dysregulated enhancers detected enriched and depleted mCA levels at MeCP2-repressed and MeCP2-activated enhancers respectively, while mCG levels showed limited variation at these regions (Figure 2.4C; 2.12B).

Given the high affinity of MeCP2 for mCG (Meehan et al., 1989), but limited mCG/CG signal at MeCP2-regulated enhancers, we further considered if the number of mCG sites rather than the level of mCG contributes to enhancer repression by MeCP2. The number of mC binding sites for MeCP2 at enhancers is determined both by per-base methylation at CA and CG sites (e.g. mCG/CG) and dinucleotide frequencies at each enhancer (e.g. CG/kb). This is particularly relevant for CG dinucleotides, which are depleted from the genome and non-uniformly distributed (Bird, 1980). We therefore quantified mC "density" (mC/kb) at dysregulated enhancers and detected a robust signal for mCG density, which is driven by the presence or absence of CG dinucleotides (Figure 2.4C; 2.12B). These findings suggest that both mCA and mCG do contribute to MeCP2 enhancer regulation, but mCA enrichment is driven by per-base methylation, while mCG enrichment is driven by CG frequencies.

To further explore the link between high-affinity sites for MeCP2 and enhancer regulation, we examined enhancer effects at higher stringency. While mCA is the highest affinity non-CG dinucleotide for MeCP2 (Gabel et al., 2015), the third nucleotide also affects MeCP2 binding, with mCAC showing highest affinity for MeCP2 (Lagger et al., 2017). We find that

mCAC is most significantly associated with MeCP2-regulated enhancers compared to lower affinity sites (Figure 2.4D; 2.12C). Consistent with the presence of high-affinity MeCP2 sites at dysregulated enhancers, MeCP2 ChIP signal is enriched at MeCP2-repressed enhancers (Figure 2.4E). Analysis of high-stringency MeCP2-repressed enhancers, selected based on enriched MeCP2 binding, yields similar changes in H3K27ac and mC enrichment (Figure 2.12A,B). This supports a direct role for MeCP2 in enhancer regulation.

While significantly dysregulated enhancers display robust signal for mCA and mCG, methylation at enhancers occurs in a continuous distribution genome-wide (Figure 2.12D). Thus, every enhancer may be regulated by MeCP2 to some degree, with the number of mC sites determining the repression level. Indeed, when we compared mC density and H3K27ac changes at all enhancers in the genome, we detected a positive correlation between mCA and mCG density and H3K27ac changes in the MeCP2 KO and a negative correlation in the MeCP2 OE (Figure 2.4F,G). While mC density is enriched in and around MeCP2-repressed enhancers (Figure 2.4C), we find that enhancer H3K27ac dysregulation in MeCP2 mutants best correlates with mC density at enhancer centers (Figure 2.4H,I). Thus, mC specifically within the enhancer is most important for regulation by MeCP2. Together, these findings support a model in which repression of enhancers by MeCP2 genome-wide is proportional to the mC density at each enhancer.

We also noted minor increases of H3K27ac at sites outside of stringently defined enhancer regions in the MeCP2 KO (Figure 2.12E). Examination of a set of possible regulatory elements (ATAC-seq peaks), defined across 13 mouse tissues (Cusanovich et al., 2018), showed that these regions overlap with enhancers as well as sites of H3K27ac enrichment not formally

called as enhancers (Figure 2.12E,F, see methods). Analysis of these regions genome-wide revealed noisier, but detectable correlations between H3K27ac changes and mC density at these putative regulatory sites (Figure 2.12F,G,H). Thus, while a larger absolute change in H3K27ac is observed at called enhancers in MeCP2 mutants (Figure 2.12F), altered H3K27ac outside of enhancer regions may also contribute to effects observed in these mice.

We next tested the role of mCA and mCG in enhancer repression by selectively disrupting mCA in neurons. Conditional *Dnmt3a* deletion in the brain ablates mCA while preserving mCG, and leads to overlapping, but smaller effects on MeCP2-repressed genes compared to MeCP2 mutants (Gabel et al., 2015). We speculated that these partial expression effects arise due to loss of enhancer repression by mCA, but preserved repression by mCG. To test this, we deleted *Dnmt3a* from post-mitotic neurons in a DNMT3A Baf53b-cKO mouse (see methods) and measured effects on mC, RNA, and H3K27ac. Bisulfite-seq revealed ablation of mCA, but retention of mCG, at enhancers and other genomic elements (Figure 2.5A, 2.13A-C), and RNA-seq showed significant, but partial, dysregulation of MeCP2-regulated genes (Figure 2.5B). H3K27ac ChIP-seq similarly detected robust, but partial, dysregulation of MeCP2-regulated enhancers (Figure 2.5B). Significantly up- and downregulated enhancers defined in the DNMT3A Baf53b-cKO (see methods) displayed enrichment or depletion of mCA, but equivalent mCG (Figure 2.5C, 2.13D,E). Unthresholded analysis showed that, like MeCP2 mutants, H3K27ac changes correlate with wild-type mCA density at enhancers genome-wide (Figure 2.5D). In contrast, no association was detected for mCG (Figure 2.5D). Together, the robust but smaller effects on MeCP2-regulated enhancers and lack of mCG-associated changes

in the DNMT3A mutant demonstrate the role of mCA in enhancer repression by MeCP2 and also support a function for mCG in this regulation.

Enhancer repression by MeCP2 is associated with MeCP2-mediated gene regulation.

We next investigated if enhancer dysregulation is linked to altered gene expression in MeCP2 mutants. As predicted by our analysis above (Figure 2.2E), enhancers in high-mCA TADs containing MeCP2-repressed genes are enriched for mCA (Figure 2.6A), suggesting possible repression by MeCP2. Analysis of gene-enhancer links by several methods (TAD overlap, GREAT analysis (McLean et al., 2010), and Hi-C) revealed that MeCP2-repressed enhancers are significantly associated with MeCP2-repressed genes (Figure 2.6B). In light of the genome-wide association between mC density and repression of H3K27ac at enhancers (Figure 2.4F,G), we considered that many high-mCA enhancers in TADs with MeCP2-repressed genes might undergo upregulation in the MeCP2 KO that are below the statistical significance cutoff. Supporting this idea, H3K27ac changes for enhancers in TADs with prominent MeCP2-repressed genes displayed trends toward upregulation (Figure 2.6C). Indeed, the population of enhancers found in TADs containing MeCP2-repressed genes showed significant upregulation as a group in the MeCP2 KO and downregulation in the MeCP2 OE (Figure 2.6D).

While this analysis linked MeCP2 enhancer repression to gene regulation, we sought to understand our previous finding that MeCP2 mutant mRNA changes are better correlated with mC levels in long genes than with mC levels in gene-flanking regions (Figure 2.8F-H; 2.11E,F) (Kinde et al., 2016). We speculated this signal might reflect preferential repression of intragenic enhancers by MeCP2. Indeed, H3K27ac changes at prominent MeCP2-repressed genes in the

MeCP2 KO showed stronger increases at intragenic enhancers than at extragenic enhancers (Figure 2.6C). Analysis of all MeCP2-repressed enhancers showed they tend to be intragenic (Figure 2.6E), and significantly overlap with MeCP2-repressed genes (Figure 2.6B). As a population, enhancers in MeCP2-repressed genes are more robustly upregulated in the MeCP2 KO and downregulated in the MeCP2 OE than all enhancers in the same TAD (Figure 2.6D). Furthermore, while H3K27ac repression by MeCP2 is correlated with mC density for both intragenic and extragenic enhancers, for a given mC density, intragenic enhancers are more repressed (Figure 2.6F).

Given the susceptibility of intragenic enhancers to MeCP2 repression, we investigated their interactions with promoters of target genes. Hi-C data revealed that intragenic sequences interact more with promoters than extragenic sequences do (Figure 2.6G), and intragenic enhancers contact promoters of their cognate gene more than equidistant extragenic enhancers do (Figure 2.6H). Quantitative 3C analysis of the MeCP2 KO detected no evidence of altered interaction strength between enhancers and promoters of MeCP2-repressed genes (Figure 2.12I). This suggests that intragenic enhancers are privileged over extragenic enhancers to control their cognate genes, but that MeCP2 does not regulate looping between enhancers and promoters. Rather, MeCP2 may affect the degree of gene activation driven by enhancers once they interact with their target promoter.

If the repression of genes containing high gene body mCA is largely due to repression of enhancers in these genes, we reasoned that gene body mCA and gene upregulation in the MeCP2 KO should only be tightly linked for genes that contain enhancers and not be linked for genes that do not contain enhancers. Because long genes contain more enhancers than shorter genes

(Figure 2.12J), MeCP2 intragenic enhancer repression could also explain the dysregulation of longer genes observed in MeCP2 mutants (Figure 2.8F-H) (Gabel et al., 2015; Sugino et al., 2014). We therefore examined gene expression effects in MeCP2 mutants for genes that do or do not contain intragenic enhancers. We detect little correlation between gene dysregulation and gene body mCA density for genes that do not contain enhancers but a robust correlation for genes containing intragenic enhancers (Figure 2.6I, 2.12L). Importantly, we observe these effects when controlling for gene length (Figure 2.12K). Consistent with the idea that enhancers specifically contribute to this effect, we find that H3K27ac changes in intragenic enhancers are more predictive of the gene expression changes in MeCP2 mutants than H3K27ac changes in non-peak sequences (Figure 2.12M). These results suggest that dysregulation of intragenic enhancers contributes to the dysregulation of long, highly-methylated genes that we previously observed in MeCP2 mutants.

2.4 Discussion

Our findings suggest a model in which TADs shape DNMT3A activity across the neuronal genome during early postnatal development, establishing mCA set-points for genes and enhancers within each TAD (Figure 2.7A). In mature neurons, MeCP2 reads-out mC to repress enhancer elements and control gene expression. Because MeCP2 most potently represses highly-methylated, intragenic enhancers, it has the largest impact on genes in high mCA TADs that contain multiple enhancers (Figure 2.7B).

The enriched DNMT3A binding and resulting high-mCA set-point we detect in TADs with MeCP2-repressed genes provides insight into the origin of high mCA previously noted at

these genes (Kinde et al., 2016; Lagger et al., 2017). The large fluctuations in mCA observed between TADs can occur because of the dynamic distribution of mCA across the genome. This contrasts with consistently high mCG across large regions. While sequence and activity states of genes and enhancers are known to affect their methylation status (Schübeler, 2015), we show here that TAD mCA set-points also influence mCA levels at genes and enhancers. Enhancers in high-mCA TADs are enriched for mCA and are more repressed by MeCP2 compared to enhancers in low-mCA TADs. In this way, TAD mCA patterns set the stage for MeCP2-mediated gene regulation in the brain.

Previously, we and others have proposed that binding of MeCP2 in genes might block RNA polymerase processivity to downregulate expression (Cholewa-Waclaw et al., 2019; Kinde et al., 2016). Surprisingly, we did not detect disrupted processivity in the MeCP2 mutants, but instead observed changes in promoter activity. Consistent with this finding, a parallel study used multiple methodologies to identify regulation of transcription initiation by MeCP2 (Boxer et al., co-submitted). Our identification of intragenic enhancers that are dysregulated in MeCP2 mutants and analysis linking them to genes (Figure 2.6I, 2.12J-M) supports the model in which MeCP2 preferentially represses long, highly-methylated genes but points to enhancer repression as a mediator of this regulation.

MeCP2 binds nearly ubiquitously across the genome and its relevant sites of action have not been clear. Through genome-wide H3K27ac analysis, we uncovered a role for MeCP2 in enhancer repression that is driven in part by TAD-scale fluctuations in levels of mCA. While identification of enhancers significantly changed in MeCP2 mutants revealed critical determinants for MeCP2 regulation (e.g. mCA and mCG density), our unthresholded analysis

suggests that these enhancers represent only extreme examples along a continuum, in which graded repression by MeCP2 that is proportional to mC density occurs at enhancers genome wide. Alterations in pervasive tuning of enhancers may explain why disruption of MeCP2 leads to profound nervous system deficits, despite the small magnitude of effects observed for individual genes.

Consistent with some role for mC sites in recruiting MeCP2 to the genome, MeCP2 ChIP-seq shows a measurable association with mCA and mCG sites (Figure 2.2D, 2.10I). However, the binding of the protein is very broad, and signal is also present at demethylated sites. Thus, mC sites may modulate MeCP2 repression of enhancers when it is bound, rather than being a strict determinant of binding. Future analyses will be necessary to dissect the role of mC sites in MeCP2 recruitment versus modulation of its repressive activity.

(Ebert et al., 2013; Kokura et al., 2001; Lyst et al., 2013) Our findings suggest a model in which MeCP2 at enhancers can induce histone deacetylation to block their activating effects on target promoters. This could occur through the interaction between MeCP2 and the NCoR histone deacetylase complex that is critical for MeCP2-mediated gene repression (Kokura et al., 2001; Lyst et al., 2013; Nott et al., 2016). Alternatively, MeCP2 could modulate NCoR functions on non-histone proteins to affect enhancer activity.

A striking finding from our study is that MeCP2 represses intragenic enhancers more than extragenic enhancers, and intragenic enhancer dysregulation is closely linked to gene changes in MeCP2 mutants. We observed that intragenic enhancers more readily contact their cognate promoters than extragenic enhancers, suggesting intragenic enhancer effects in MeCP2 mutants are particularly impactful for gene expression. Notably, intragenic enhancer regulation

by MeCP2 provides an explanation for the gene length and gene body mC signature that we previously observed for MeCP2-repressed genes (Kinde et al., 2016). These findings also suggest that the presence of an enhancer within a gene can fundamentally alter the nature of its regulation. Future studies can dissect how MeCP2 preferentially represses intragenic enhancers, providing insight into the unique regulatory environment of these elements.

Our study focused on mCA at MeCP2-repressed genes, but enhancer repression by MeCP2 may also explain MeCP2-activated genes. The MeCP2-activated enhancers we identify here are depleted of mC sites (Figure 2.4C; 2.12B-D). If MeCP2 primarily represses enhancers through mC, then H3K27ac changes at "activated" enhancers could reflect a relative lack of de-repression in the MeCP2 KO and escape from repression in the MeCP2 OE. MeCP2-activated enhancers are associated with MeCP2-activated genes (Figure 2.6B), suggesting the relative changes in gene expression may stem from enhancer effects. Future studies can test if a single model of enhancer repression by MeCP2 applies, or if MeCP2 directly activates genes by other mechanisms.

While robust changes occur at enhancers in MeCP2 mutants, the protein binds broadly across the genome and disruption of MeCP2 has effects on acetylation outside of stringently defined enhancer sequences (Figure 2.12E) (Boxer et al., co-submitted). While some of this signal appears to originate at subthreshold enhancers (Figure 2.12E), MeCP2 is likely to have repressive effects on chromatin outside of enhancers that can contribute to gene regulation.

Multiple datasets support a role for MeCP2 in repressing long, high-mCA genes (Gabel et al., 2015; Rube et al., 2016; Sugino et al., 2014), but a recent study proposed that long gene upregulation in MeCP2 mutants results from post-transcriptional rather than transcriptional

effects (Johnson et al., 2017). Another study suggested long-gene dysregulation in individual MeCP2 datasets does not reach statistical significance (Raman et al., 2018). Here, independent RNA-seq and ChIP-seq analyses further show that MeCP2 preferentially affects long, high-mCA genes. Notably, a parallel study using multiple methodologies and large-replicate datasets has also verified these effects (Boxer et al., co-submitted). Our additional insights into enhancer regulation not only support these effects but also provide a mechanism for how they can occur.

Our findings have important implications for Rett syndrome and MeCP2 duplication disorder. Mutation of MeCP2 in these disorders is likely to disrupt enhancer activity in high-mCA TADs, driving altered gene expression, and contributing to disease pathology. In this way, our study suggests that Rett syndrome and MeCP2 duplication syndrome are disorders that stem in part from disruption of enhancer control.

2.5 Methods

Mice

MeCP2 knockout mice (B6.129P2(C)-*MeCP2*$^{tm1.Bird}$/J) were obtained from The Jackson Laboratory. Female heterozygous mice (*MeCP2*$^{-/+}$) were crossed with C57BL/6J male mice to generate hemizygous male knockout mice (*MeCP2*$^{-/y}$) and wild-type male litter mates (*MeCP2*$^{+/y}$). *MeCP2* overexpression mice (FVB-Tg(MECP2)3Hzo/J) were cryo-recovered from The Jackson Laboratory. Female heterozygous mice (*MeCP2*$^{Tg3/+}$) were crossed with FVB/NJ male mice to generate hemizygous male transgenic mice (*MeCP2*$^{Tg3/y}$) and wild-type male litter mates (*MeCP2*$^{+/y}$). Female *Dnmt3a*$^{fl/fl}$ were provided by M. Goodell and crossed to male B6.Cg-Tg(Nes-cre)1Kln/J (*Nestin*-Cre$^{+/-}$) to generate *Dnmt3a*$^{fl/+}$; *Nestin-Cre*$^{+/-}$. Male *Dnmt3a*$^{fl/+}$; *Nestin-Cre*$^{+/-}$ were then crossed to female *Dnmt3a*$^{fl/fl}$ to generate *Dnmt3a*$^{fl/fl}$ Tg(Nes-cre)1Kln/J conditional knockout mice (DNMT3A Nestin-cKO) (Gabel et al., 2015). To generate *Dnmt3a*$^{fl/fl}$; Tg(Actl6b-cre)4092Jiwu/J conditional knockout mice (DNMT3A Baf53b-cKO), *Dnmt3a*$^{fl/fl}$ were crossed to Tg(Actl6b-cre)4092Jiwu/J (*Baf53b*-Cre$^{+/+}$) to generate *Dnmt3a*$^{fl/+}$; *Baf53b-Cre*$^{+/-}$. *Dnmt3a*$^{fl/+}$; *Baf53b-Cre*$^{+/-}$ were then crossed to *Dnmt3a*$^{fl/fl}$ producing experimental and control animals for analysis.

Topologically associated domain analysis

Topologically associating domains were called using the TADtree algorithm (Weinreb and Raphael, 2016) on interaction matrices (Dixon et al., 2012), with the following parameters: gamma (sensitivity vs specificity tradeoff) = 200, M (number of hierarchical layers) = 1, p and q (minimum scale of interaction shift) = 3 and 12 respectively, and N (maximum number of TADs detected) = 500. To allow for manual model selection, TADtree also computes TADs for the range of 1:500, in this case. We leveraged this to apply an additional filter for consistency, wherein we selected for TADs that are called in at least 30% of all runs. When we applied this filter, we found that the effects of changing gamma were moderated, and the program performed consistently at many ranges of sensitivity vs specificity. TADs defined in the cerebral cortex by this method are ~400kb on average and range in size from ~160kb to ~2mb, a scale that is similar to regions that we observed with enriched mCA levels. We defined TAD-related "contact domains" of interaction across the genome using the Arrowhead algorithm as previously described (Rao et al., 2014) at 5kb resolution from embryonic neocortex. These contact domains had a similar length distribution to TADs, with a median length of half that of TADs. These domains are established and detected before DNMT3A increases in expression and establishes mCA, and the close correlation between early TAD structure, early DNMT3A binding, and mCA patterns suggests (Figure 2.10C,D) that the activity of DNMT3A in a TAD during early postnatal development defines a long-term "set-point" for mCA across this region into adulthood. In addition, the fact that DNMT3A expression is low at 8-weeks and that 8-week DNMT3A binding, concurrent with adult TAD structure, is less correlated with mCA compared to 2-week DNMT3A binding (Figure 2.10E,F), lends more support to a stable "set-point" of mCA, rather than levels that fluctuate through adulthood.

8-week cortex Hi-C data was obtained as pre-processed observed/expected contact matrices, from http://chromosome.sdsc.edu/mouse/hi-c/download.html (Dixon et al., 2012). For analysis of neocortex Hi-C data from Bonev and colleagues (Bonev et al., 2017), raw FASTQ files were downloaded from GEO. HiC-Pro (Servant et al., 2015) was used to generate contact matrices using the mm10 mouse genome as reference. Juicer (Durand et al., 2016) was then run on the contact matrices to generate Hi-C contact matrices at 1.5kb, 5kb, and 40kb resolutions using KR normalization. Arrowhead domains were then called on the data within Juicer. For cerebellum analysis, contact domains called from 10kb-resolution Hi-C data were used (Yamada et al., 2019).

Total and nuclear RNA isolation

Cerebral cortex was dissected on ice in phosphate buffered saline from 1) MeCP2 KO and wild-type male litter mates at 7-8 weeks old, 2) MeCP2 OE and wild-type male litter mates at 7-10 weeks old, and 3) DNMT3A Baf53b-cKO and control mice at 7-8 weeks old. Total RNA

was extracted from 1/16th of a whole cortex using RLT buffer following RNeasy Micro Kit (Qiagen). Nuclear RNA was isolated following a modified version of the protocol described (Mo et al., 2015). Briefly, half of a cortex was homogenized in 0.25M sucrose, 25mM KCl, 5mM $MgCl_2$, 20mM Tricine-KOH using a glass dounce homogenizer. Nuclei were isolated via centrifugation at 10,000g for 18 minutes at 4°C (Sorvall HB-4) by pelleting through a 30% iodixanol density gradient (Sigma D1556). RNA was isolated from nuclei by resuspending pellet in RLT buffer following the RNeasy Micro Kit (Qiagen).

RNA sequencing

RNA libraries were generated from 250ng total and nuclear RNA with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB) using a modified amplification protocol (37°C, 15 minutes; 98°C, 30 seconds; (98°C, 10 seconds; 65°C, 30 seconds; 72°C, 30 seconds)x13; 72°C, 5 minutes; 4°C hold. RNA libraries were pooled at a final concentration of 8-10nM and sequenced using Illumina HiSeq 2500 or 3000 with the Genome Technology Access Center (GTAC) at Washington University in St. Louis, typically yielding 20-30 million single-end reads per sample.

RNA sequencing analysis

Raw FASTQ files were trimmed with Trim Galore, using a quality filter of 20, then rRNA sequences were filtered out using Bowtie, using rRNA sequences from Mus Musculus obtained from the NCBI sequence database. The unaligned reads from this step were then aligned to mm9 using STAR (Dobin et al., 2013) with the default parameters. Reads mapping to multiple regions in the genome were then filtered out, and uniquely mapping reads were converted to BED files. Intronic and exonic reads were then separated. To do this, splice-site reads were first filtered out of the BED read files, then reads that mapped entirely within exons were added to the splice-site reads to make the exonic read file. All remaining reads that overlapped introns were considered intronic reads. Finally, reads were assigned to genes using bedtools coverage -counts.

For gene annotation we defined a "flattened" list of longest transcript forms for each gene, generated on Ensgene annotations, obtained from the UCSC table browser. For each gene, Ensembl IDs were matched to MGI gene names. Then, for each unique MGI gene name, the most upstream Ensgene TSS and the most downstream TES were taken as that gene's start and stop. Based on these Ensembl gene models, we defined TSS regions and gene bodies.

Chromatin immunoprecipitation protocol

Cerebral cortex was dissected on ice in phosphate buffered saline from 1) MeCP2 KO and wild-type male litter mates at 7-8 weeks old, 2) MeCP2 OE and wild-type male litter mates at 7-10 weeks old, 3) DNMT3A Baf53b-cKO and control mice at 7-8 weeks old. The tissue was

flash-frozen in liquid nitrogen and stored at -80°C. ChIP experiments were performed on half a cortex as previously described (Cohen et al., 2011), using an alternative chromatin fragmentation method. Chromatin were fragmented with Covaris E220 sonicator (5% Duty Factory, 140 Peak Incidence Power, 200 cycles per burst, milliTUBE 1mL AFA Fiber). ChIP was performed with H3K27ac (0.025-0.1µg; Abcam ab4729), H3K4me3 (2µg; Abcam ab1012), and H3K36me3 (0.2µg; Active Motif 61101). ChIP libraries for H3K27ac, H3K4me3, and H3K36me3 were generated using Ovation Ultralow Library System V2 (NuGEN). Libraries were pooled to a final concentration of 8-10nM and sequenced using Illumina HiSeq 3000 with GTAC, yielding 15-30 million single-end reads per sample.

Chromatin immunoprecipitation analysis

Sequenced reads were mapped to the mm9 genome using bowtie2 alignment, and reads were extended based on library sizes and deduplicated to consolidate PCR duplicate reads. Deduplicated reads were used to quantify read density normalized by the number of reads per sample and by read length in basepairs. Bedtools coverage –counts was used to quantify ChIP signal at the transcriptional start site (TSS), gene body (GB), and transcriptional end site (TES). For consistency with methylation analysis, the TSS was defined as a 1kb region surrounding the TSS (+/-500bp), the GB was defined as 3kb downstream of the TSS to the end of the transcript, and the TES was defined as 2kb upstream through 3kb downstream of the end of the transcript, based on our Ensembl gene models. edgeR was then used to determine differential ChIP-signal across genotypes.

Whole-genome bisulfite sequencing

Cerebral cortex was dissected from DNMT3A Nestin-cKO or DNMT3A Baf53b-cKO and controls at 7-8 weeks, flash-frozen in liquid nitrogen, and stored at -80°C. DNA extraction and bisulfite conversion were performed as before (Gabel et al., 2015). Briefly, genomic DNA was extracted and bisulfite libraries were generated using the Ovation Ultralow Methyl-Seq Library System (NuGEN). Libraries were pooled and sequenced using Illumina MiSeq 2x150 with the Spike-In Cooperative at Washington University in St. Louis.

Chromatin conformation capture

3C assays were adapted from previously described procedures (Kim and Dekker, 2018; Lieberman-Aiden et al., 2009; Yamada et al., 2019). Half of a cerebral cortex was dounced 10x with a loose pestle in cross-linking buffer (50mM Hepes-KOH, pH 7.9; 100mM NaCl; 1mM EDTA; 0.5mM EGTA) with 1% formaldehyde at room temperature for 15 minutes. Formaldehyde was quenched with 125mM glycine and incubated for 5 minutes at room temperature. The suspension was pelleted at 1150xg for 5 minutes at 4°C and the pellet was subsequently washed with PBS and frozen at -80°C. The frozen pellet was thawed and lysed

(10mM Tris-HCl, pH 8; 10mM NaCl, 0.2% IGEPAL-630, Protease Inhibitor (Complete Protease Inhibitor Cocktail, Roche)) for 15 minutes followed by douncing 25x with a loose pestle on ice. The lysed cells were strained through a 70um cell strainer (Falcon) and pelleted at 200xg for 5 minutes at 4°C. Nuclei were resuspended in 1mL of 0.5% SDS and incubated at 62°C for 10 minutes to permeabilize the nuclei. 25ul (approx. 90-100k nuclei) were used for the 3C library construction, and SDS was quenched with Triton X-100 (1% Triton, 1x NEB 3.1 Buffer). Nuclei were digested overnight at 37°C with 200U of BglII. The digested nuclei were incubated at 65°C for 15 minutes to inactivate the enzyme. Samples were then ligated for 6 hours at 16°C with 4000U of T4 DNA ligase (NEB Ligase Buffer, 0.1mg/ml BSA, Triton X-100, NEB T4 ligase). Digested and ligated 3C samples were pelleted at 3500rpm for 5 minutes at 4°C and resuspended in 200ul NEB 3.1 Buffer and brought to 1% SDS and 250mM NaCl. 3C libraries were de-crosslinked overnight at 65°C. They were then incubated with 40ug RNase A at 37°C for 1 hour, followed by 80ug of Proteinase K at 55°C for 1 hour. The samples were purified with Phenol/Chloroform/Isoamyl Alcohol (25:24:1) and subsequently ethanol precipitated (100mM Sodium Acetate) overnight at -20°C. Control libraries were generated from Bacterial Artificial Chromosomes (BACs) as previously published (Kim and Dekker, 2018). BACs (see STAR methods) were isolated using PureLink HiPure Plasmid Maxiprep Kit (Invitrogen), following the modified protocol for BAC isolation.

Quantification and Statistical Analysis

Whole-genome bisulfite analysis

Bisulfite-sequencing cannot distinguish between hydroxymethylation and methylation at cytosines, detecting both only as modified sites during sequencing. Thus, measures of methylation included in this study represent the aggregate of both forms of methylation at sites across the genome. Bisulfite data for the DNMT3A Nestin-cKO and DNMT3A Baf53b-cKO, and data obtained from GEO as FASTQ files, were adapter-trimmed, mapped to mm9, then deduplicated and called for methylation using BS-seeker2 (with bowtie2). Nonconversion rate was set to be (--XS=.3,2), and default settings were used otherwise. The methylation levels for genes and regions were assessed by summing the number of reads mapping to Cs that supported mC and dividing that by the number of reads mapping to Cs that supported non-mC, using bedtools map -o sum. This allows sites with more read information to contribute more in determining the methylation level of the surrounding region. In order to avoid confounding effects of promoter-associated depletion of methylation, genic methylation was assessed 3kb downstream of promoters to the TES. In order to assess any potential for C site coverage bias influencing TAD-associated Bisulfite-seq results, a bed file of every C/G in the genome was generated, and bedtools coverage -counts was run on it. Bisulfite coverage bias was assessed on informative sites for methylation with bedtools map -o mean on the number of reads mapping to Cs in the region.

RNA-sequencing quantification

MeCP2-repressed and activated genes were identified by quantitative analysis of exonic reads from total RNA from the MeCP2 KO and OE. For this analysis we applied an approach similar to previous studies of other brain regions that combined results of gene expression in these two strains in order to identify the most-robustly MeCP2-regulated genes (Ben-Shachar et al., 2009; Chahrour et al., 2008; Chen et al., 2015). DESeq2 was run using default parameters on exonic reads from MeCP2 KO and their littermate control animals (n=6 per genotype). Separately, we ran DESeq2 analysis on exonic reads from MeCP2 OE and their littermate control animals (n=5). The nominal p-values output by DESeq2 for each gene in each mutant-control comparison were then combined using the Fisher method (log-sum). The resulting combined p-values were then Benjamini-Hochberg corrected, and genes with a q-value < .1 and a log2 fold-change > 0 in the KO and a log2 fold-change < 0 in the OE were labeled as MeCP2-repressed, while genes with a q-value < .1 and a log2 fold-change < 0 in the KO and a log2 fold-change > 0 in the OE were labeled as MeCP2-activated. Notably, similar results for enrichment of mCA, and gene length were observed when examining lists of genes called as significantly dysregulated in the MeCP2 KO and OE on their own. However, a smaller gene list was identified in each independent analysis, likely due to the reduced statistical power. For comparison between changes in intronic RNA and changes in exonic RNA (Figure 2.3; S1) a list of significantly changed genes in the MeCP2 KO compared to littermate controls based on intronic fold-changes was generated by applying the same DESeq2 analysis to intronic reads derived from total RNA-seq of nuclear RNA.

RNA-seq aggregate plots examining changes in expression over lengths of genes (Figure 2.11G-I) were performed by binning genes into 1kb windows, then calculating nuclear intronic coverage over each bin using bedtools coverage -hist. Each gene was then normalized by the median amount of combined coverage from the MeCP2 KO and wild-type. Finally, genes were aligned by their TSSs, and median expression levels were plotted for each bin. For any given graph, the genes are filtered such that the lengths of genes plotted is equal to or greater than the aggregate length being plotted (Boswell et al., 2017; Gray et al., 2014). Termination ratios (adapted from (Boswell et al., 2017)) were calculated in a manner similar to the aggregate plots of expression. Windows were made for every gene longer than 100kb, from the TSS to 25kb downstream of the TSS (window A), and from the TES to 25kb upstream of the TES (window B). Coverage over these windows was calculated with bedtools coverage -hist, then coverage of window A was divided by the coverage of window B, within each sample. Then, wild type/MeCP2 KO samples were divided against each other, and a normalized metric for each wild type/MeCP2 KO replicate was generated. Finally, median metrics were plotted for each wild

type/MeCP2 KO replicate. Simulated metrics were generated by progressively applying exonic fold-changes to the wild-type expression throughout the gene. For example, a gene 50kb long, with an observed exonic log2 fold-change (MeCP2 KO/wild type) of -1 would be divided into 50 1kb bins. A simulated KO expression pattern would then be generated by progressing down each of the 50 bins, applying a fold-change of 1/50 * -1, 2/50 * -1, etc, to the expression of the corresponding bin in the WT.

Controlled resampling

A resampling approach that controls for a desired variable was used throughout the paper. A sample set (e.g. MeCP2-repressed genes) and a control set (e.g. all other genes) are assessed for a certain characteristic (e.g. expression, length, etc). The control set is then sorted on this characteristic, and each entry in the sample set is assessed for where it would be placed in the sorted control set. Then, for each sample entry, a control entry is selected that is within 10 places of the sample entry, generating a control set the same size and variable distribution as the sample set.

TAD boundary analysis

For Figures 2.1F, 2.2C, 2.9B, and 2.9F, all TADs that intersected MeCP2-repressed, MeCP2-activated, and unchanged genes were selected, boundaries phased, and values for mC/C, coverage, GC percent, or ChIP/Input plotted in aggregate using R, python, and bedtools. Shuffling analysis was performed in these cases by moving each TAD randomly around their target genes (TADs containing MeCP2-repressed genes were shuffled around MeCP2-repressed genes, etc.). This was done once to generate a single example (e.g. Figure 2.1F), or up to 20 times to generate a resampling ribbon (e.g. Figure 2.9B). Figures 2.1G and 2.9K were generated by selecting boundaries through the following method: TADs were assigned a score, based on the difference between its mCA level and the mCA level of the next downstream TAD. Then, the 33% top scoring TADs were taken and plotted in aggregate analysis. For shuffled TAD control plots, TAD locations were randomized and the selection process for these shuffled TADs was repeated. We note that because both true TADs and shuffled TADs in these plots were selected based on high differences in average methylation between them, a reduction of signal going from left to right is predicted. However, true TADs display two aspects of mCA/CA signal that are not present in shuffled TADs, which demonstrate the organizing effects associated with TAD boundaries: 1) True TADs are more highly enriched for mCA, showing that resampling TADs eliminates an organized enrichment of mCA/CA in specific regions of the genome. 2) A steep step-down in mCA/CA occurs at true TAD boundaries, while a more gradual fall off in signal occurs at reshuffled boundaries. These two differences illustrate that ~30% shifts in mCA/CA levels are common between adjacent TADs, and they demonstrate that TAD boundaries delineate a sharp transition between mCA/CA levels at these regions. Plotted in all figures is the

average of the upstream and downstream TAD boundaries, taking into account TAD orientation. Upstream boundaries (where the TAD would be on the right, instead of the left), have their orientation flipped horizontally along the boundary. The value plotted is the mean of the flipped upstream boundaries and the unflipped downstream boundaries.

TAD methylation cross-correlation

Cross-correlation matrices (e.g. Figure 2.1H) were generated by dividing each domain into 10 equally-sized bins, then prepending and appending 10 identically sized bins up and down-stream of the domain, making a number of domains x 30 matrix. Each column of this matrix was then correlated against each other, making a 30 x 30 correlation matrix, which was plotted in heatmap form. Shuffled TADs were generated as a negative control by randomly placing TAD-sized regions around the genome, separated from each other by similar distances as actual TADs to retain TAD structure.

In order to calculate heatmaps of correlations between mCA/CA levels of genomic elements in and outside of TADs and mean TAD mCA/CA levels (Figure 2.10J), TADs were again divided into 10 equally sized bins, with equal-sized regions placed upstream and downstream the domain. Enhancers, genes, and TSSs in the genome were then intersected with these regions, and Spearman correlations between TAD methylation (subtracting out the methylation of the element if necessary) and element methylation within each region were calculated.

To assess the similarity of mCA levels between individual elements inside and outside of the same TAD (i.e. TSS regions, enhancers, gene bodies; Figure 2.10K), each element was paired to each other element on the same chromosome. Each pair was then assessed if they paired within or between TADs. Because mCA varies with genomic distance, each intra-TAD pair was matched to the extra-TAD pair with the most similar distance between elements, and Spearman correlations were calculated on the two distance-matched sets.

Identification of enhancers

Enhancers in this study were defined by stringent criteria requiring the presence of overlapping H3K27ac and H3K4me1 peaks that occur outside of a known TSS region, or a peak of the promoter-associated histone mark H3K4me3. As noted, this led to the exclusion of some subthreshold regions of H3K27ac enrichment that may represent true regulatory elements, but ensured that we analyzed robust regulatory elements in our studies. For this analysis, bed files of H3K27ac and H3K4me3 ChIP-seq were pooled by replicate. Peaks of H3K27ac and H3K4me3 ChIP-seq were identified using the MACS2 peak calling algorithm on the pooled bed files using the pooled ChIP Input as background signal (macs2 callpeak --nomodel -q 0.05). MeCP2 KO and wild type peak files were then combined using bedtools unionbedg, and overlapping peaks were merged into single peaks using bedtools merge. Bedtools intersect was used to identify

H3K27ac peaks that did not overlap with gene promoter regions (1kb around annotated TSS) or with H3K4me3 peaks from MACS2. These non-overlapping H3K27ac peaks were then further filtered for landing within an H3K4me1 peak, as called in the ENCODE-generated broadpeak file for H3K4me1 ChIP from 8-week old cortex. All H3K27ac peaks that remained after these rounds of filtering were defined as enhancers.

To identify the enhancers most robustly regulated by MeCP2, we used combined analysis of the MeCP2 KO and OE, similar to the approach used for determining gene-expression changes. We ran differential ChIP-seq analysis on H3K27ac from the MeCP2 KO and their littermate control animals (n=5), and from the MeCP2 OE and their littermate controls (n=3). Reads were quantified in all merged acetyl peaks, and edgeR was used to calculate nominal p-values and fold-changes for these peaks. These p-values were then combined using the Fisher method (log-sum) and were Benjamini-Hochberg corrected. Acetyl peaks with a combined q-value < 0.1, and a log2 fold-change > 0 in the KO and a log2 fold-change < 0 in the OE were called as MeCP2-repressed peaks, while peaks with a combined q-value < 0.1, and a log2 fold-change < 0 in the KO and a log2 fold-change > 0 in the OE were called as MeCP2-activated peaks.

Enhancers were also called as misregulated in the DNMT3A Baf53b-cKO, using edgeR. H3K27ac ChIP-seq reads from the DNMT3A Baf53b-cKO and control (n=6) were quantified in the merged acetyl peaks called from the MeCP2 KO/control. edgeR was then run on these regions, and peaks with a q-value < 0.1 were called as misregulated in the DNMT3A Baf53b-cKO.

ATAC peak analysis

To sensitively detect sub-peak-threshold histone acetylation signal that could correspond to putative regulatory elements, a compendium of all ATAC peaks detected in the genome was obtained from http://atlas.gs.washington.edu/mouse-atac/data/ (atac_matrix.binary.qc_filtered.peaks.txt) and acetylation was quantified in these regions. ATAC peaks that landed within 1kb of a TSS were filtered out, and methylation and acetylation analysis was performed on them as described for enhancers. Equal-sized control regions for enhancers and ATAC peaks were generated through a structured resampling approach. Enhancers and ATAC peaks, if located within a gene, are shuffled within that gene. If the peak is extragenic, it is shuffled between the nearest upstream and downstream enhancer/gene/ATAC peak. In each case, enhancers and ATAC peaks are restricted from landing within existing enhancers and ATAC peaks. Normalized pseudocounts of acetylation for enhancers, ATAC peaks, and resampled ATAC peaks were generated from edgeR common dispersions, running on each dataset and mouse strain separately.

Associations between enhancers and genes using Hi-C and GREAT

GREAT 3.0 (McLean et al., 2010) was used to determine potential enhancer-promoter interactions. NCBI build 37 of Mus Musculus was used as the species assembly, and bed files of enhancer regions were uploaded into the web tool. Enhancer-promoter interactions were identified for MeCP2-repressed, MeCP2-activated, and all other enhancers. All identified genes linked to these enhancers were used for further analysis. To link enhancers to promoters by Hi-C, intrachromosomal Hi-C matrices (KR-normalized) were extracted at 1.5kb resolution using Straw (Durand et al., 2016), and mean average interactions for all distances were calculated per-chromosome. Each matrix was then filtered down to only interactions between promoters and enhancers within 3mb of each other, and enhancers-promoter pairs with greater than 3 observed interactions and an observed/expected ratio over 1.5 were linked, making Hi-C-linked enhancers. Enhancers that were linked to promoters of MeCP2-repressed or MeCP2-activated genes were then used in analysis in Figure 2.6B.

For analysis of intragenic versus extragenic contact frequencies (Figure 2.6G), the same normalized contact matrices were analyzed. Genes greater than 50kb were extended on either side by their respective gene lengths, and contacts were mapped to them using bedtools intersect -wao -F 1. Each region was then split into 60 equally-sized bins (20 upstream, 20 intragenic, 20 downstream), making 3600 possible regions of interactions. Average interaction frequencies within these regions were calculated from the interaction matrix, and each gene's intragenic and extragenic interactions were aggregated by calculating the mean of each bin of interaction.

For Figure 2.6H, intragenic and extragenic enhancer-promoter interactions were distance matched to control for the greater number and variability of extragenic interactions. To do this, a similar resampling approach to Figure 2.10K was used: for each intragenic interaction, an extragenic interaction with a similar distance was selected for comparison.

To assess the relationship between enhancer acetylation fold-change and gene fold-change (Figure 2.12M), partial correlations between enhancer/control region acetyl fold-change and gene fold-change were calculated by averaging the acetyl fold-changes of elements within the gene and correlating this aggregate value to that gene's fold-change.

Quantitative 3C analysis

Quantitative PCR was employed, as previously described (Joo et al., 2016; Schaukowitch et al., 2014; Tolhuis et al., 2002), to determine enhancer-promoter interactions of enhancers using the primers listed. Negative regions were selected from genomic restriction fragments that did not contain detectable enhancer sequences and were located nearby to the enhancer being tested but in closer proximity to the anchor TSS. Relative concentrations of enhancers and corresponding negative regions were calculated from a standard curve of BAC 3C libraries containing the targeted loci (see STAR methods). Enrichments of interactions were then calculated as the relative concentration of targeted enhancers divided by the relative concentration of the nearby negative region.

mC context enrichment analysis

Trinucleotide contexts and methylation status (a rational number determined through bisulfite coverage - # of Cs mapping to site / # of Ts mapping to site) were determined for each cytosine and guanine in the genome, and then sites were assigned to enhancers based on proximity (a site was assigned to an enhancer if it was within 1 kb of the enhancer's center). T-tests were then run for each context, summing methyl-weighted or unweighted (for total mC per kb / C sites per kb respectively), or averaging methylation statuses (for percent mC), comparing test sets (up/downregulated enhancers, or enhancers within certain genes) to a set of resampled control enhancers with similar acetylation levels 5 times the size of the test set.

mC vs acetylation change local correlation analysis at enhancers

For both enhancers and gene local correlation analysis, 1 kb-sized bins were assessed for average methylation around and within enhancers/genes, making an N x M sized matrix of methylation, where N = # of enhancers/genes, and M = # of bins. Each column of this matrix was then correlated against the matching enhancer/gene's fold-change in the MeCP2 KO/MeCP2 OE. For enhancers, correlation analysis was centered on summits of wild-type acetylation and MeCP2 KO acetylation and analyzed separately through MACS2.

Running-average plots

Running-average plots of mCA/CA and genic fold-change as well as total mC sites and genic fold-change were generated from means of 201 gene bins, with a 1 gene step, using the rollMean command in the zoo package of R. Length-controlled resampling was performed by selecting a gene within .75 – 1.25 times the length of each gene in the test set.

TAD/element methylation correlation

Enhancers, transcriptional start sites, and gene bodies were intersected with TADs and contact domains, and each unique intersection was plotted (Figure 2.2E, 2.9L,M). Shuffling analysis was generated by moving elements randomly within its containing TAD. To avoid spuriously detecting a correlation due to the contribution of the elements themselves to the calculated average methylation of the TAD, the mC signal from the elements within the TAD (e.g. gene) were excluded from the calculation of the mC levels for that TAD.

Misregulated gene-enhancer linkage enrichment

To calculate the significance and magnitude of the linkage between misregulated genes and enhancers (Figure 2.6B), misregulated enhancers were resampled based on acetylation, and a fisher's exact test was performed comparing the association of enhancers with misregulated

genes between the misregulated and control set. This process was repeated 1000 times, and the median p-value and log2 fold-enrichment was plotted.

Data and Software Availability

All genomic data generated in this study have been uploaded to the NCBI GEO archive GSE123373

2.6 Acknowledgements

We thank Y. Liu and M. Nemera and the Gabel lab, as well as J. Dougherty, K. Kroll, J. Edwards, A. Bonni, and J. Yi for support and feedback on the manuscript. We thank L. Boxer, W. Renthal, and M. Greenberg for discussions. The following grants supported this work: NIH 5T32GM007067 and F31NS108574 to A.W.C; The Klingenstein-Simons Fellowship Fund, the G. Harold and Leila Y. Mathers Foundation, the Brain and Behavior Research Foundation, and NIMH R01MH117405 to H.W.G.

2.7 References Cited

1. Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185–188.
2. Baker, S.A., Chen, L., Wilkins, A.D., Yu, P., Lichtarge, O., and Zoghbi, H.Y. (2013). An AT-Hook Domain in MeCP2 Determines the Clinical Course of Rett Syndrome and Related Disorders. Cell *152*, 984–996.
3. Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial Distribution of Di- and Tri-methyl Lysine 36 of Histone H3 at Active Genes. J. Biol. Chem. *280*, 17732–17736.
4. Ben-Shachar, S., Chahrour, M., Thaller, C., Shaw, C.A., and Zoghbi, H.Y. (2009). Mouse models of MeCP2 disorders share gene expression changes in the cerebellum and hypothalamus. Hum. Mol. Genet. *18*, 2431–2442.
5. Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.
6. Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557–572.e24.
7. Boswell, S.A., Snavely, A., Landry, H.M., Churchman, L.S., Gray, J.M., and Springer,

M. (2017). Total RNA-seq to identify pharmacological effects on specific stages of mRNA synthesis. Nat. Chem. Biol. *13*, 501.

8. Boxer, L.D., Renthal, W., Greben, A.W., Whitwam, T., Silberfeld, A., Stroud, H., Li, E., Yang, M.G., Kinde, B., Griffith, E.C., et al. MeCP2 represses the rate of transcriptional initiation of highly methylated long genes. Co-Submitted.
9. Chahrour, M., and Zoghbi, H.Y. (2007). The Story of Rett Syndrome: From Clinic to Neurobiology. Neuron *56*, 422–437.
10. Chahrour, M., Jung, S.Y., Shaw, C., Zhou, X., Wong, S.T.C., Qin, J., and Zoghbi, H.Y. (2008). MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription. Science *320*, 1224–1229.
11. Chen, L., Chen, K., Lavery, L.A., Baker, S.A., Shaw, C.A., Li, W., and Zoghbi, H.Y. (2015). MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. Proc. Natl. Acad. Sci. U. S. A. *112*, E2982–E2982.
12. Cholewa-Waclaw, J., Bird, A., von Schimmelmann, M., Schaefer, A., Yu, H., Song, H., Madabhushi, R., and Tsai, L.-H. (2016). The Role of Epigenetic Mechanisms in the Regulation of Gene Expression in the Nervous System. J. Neurosci. *36*, 11427 LP-11434.
13. Cholewa-Waclaw, J., Shah, R., Webb, S., Chhatbar, K., Ramsahoye, B., Pusch, O., Yu, M., Greulich, P., Waclaw, B., and Bird, A.P. (2019). Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. Proc. Natl. Acad. Sci. *116*, 14995 LP-15000.
14. Cohen, S., Gabel, H.W., Hemberg, M., Hutchinson, A.N., Sadacca, L.A., Ebert, D.H., Harmin, D.A., Greenberg, R.S., Verdine, V.K., Zhou, Z., et al. (2011). Genome-Wide Activity-Dependent MeCP2 Phosphorylation Regulates Nervous System Development and Function. Neuron *72*, 72–85.
15. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. *107*, 21931–21936.
16. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.e18.
17. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
18. Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin Domains: the Unit of Chromosome Organization. Mol. Cell *62*, 668–680.
19. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,

Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

20. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. *3*, 95–98.
21. Ebert, D.H., and Greenberg, M.E. (2013). Activity-dependent neuronal signalling and autism spectrum disorder. Nature *493*, 327.
22. Ebert, D.H., Gabel, H.W., Robinson, N.D., Kastan, N.R., Hu, L.S., Cohen, S., Navarro, A.J., Lyst, M.J., Ekiert, R., Bird, A.P., et al. (2013). Activity-dependent phosphorylation of MeCP2 threonine 308 regulates interaction with NCoR. Nature *499*, 341–345.
23. ENCODE Consortium, T.E.P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature *489*, 57–74.
24. Van Esch, H., Bauters, M., Ignatius, J., Jansen, M., Raynaud, M., Hollanders, K., Lugtenberg, D., Bienvenu, T., Jensen, L.R., Gécz, J., et al. (2005). Duplication of the MECP2 Region Is a Frequent Cause of Severe Mental Retardation and Progressive Neurological Symptoms in Males. Am. J. Hum. Genet. *77*, 442–453.
25. 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 *522*, 89–93.
26. Gray, J.M., Harmin, D.A., Boswell, S.A., Cloonan, N., Mullen, T.E., Ling, J.J., Miller, N., Kuersten, S., Ma, Y.-C., McCarroll, S.A., et al. (2014). SnapShot-Seq: A Method for Extracting Genome-Wide, In Vivo mRNA Dynamics from a Single Total RNA Sample. PLoS One *9*, e89673.
27. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell *130*, 77–88.
28. 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.
29. Hagberg, B., Aicardi, J., Dias, K., and Ramos, O. (1983). A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. Ann. Neurol. *14*, 471–479.
30. 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.
31. Ip, J.P.K., Mellios, N., and Sur, M. (2018). Rett syndrome: insights into genetic, molecular and circuit mechanisms. Nat. Rev. Neurosci. *19*, 368–382.

32. Johnson, B.S., Zhao, Y.-T., Fasolino, M., Lamonica, J.M., Kim, Y.J., Georgakilas, G., Wood, K.H., Bu, D., Cui, Y., Goffin, D., et al. (2017). Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome. Nat Med *23*, 1203–1214.
33. Joo, J.-Y., Schaukowitch, K., Farbiak, L., Kilaru, G., and Kim, T.-K. (2016). Stimulus-specific combinatorial functionality of neuronal c-fos enhancers. Nat. Neurosci. *19*, 75–83.
34. Kim, T.H., and Dekker, J. (2018). Generation of Control Ligation Product Libraries for 3C Analyses. Cold Spring Harb. Protoc. *2018*, pdb.prot097865.
35. Kinde, B., Gabel, H.W., Gilbert, C.S., Griffith, E.C., and Greenberg, M.E. (2015). Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. Proc. Natl. Acad. Sci. U. S. A. *112*, 6800–6806.
36. Kinde, B., Wu, D.Y., Greenberg, M.E., and Gabel, H.W. (2016). DNA methylation in the gene body influences MeCP2-mediated gene repression. Proc. Natl. Acad. Sci. *113*, 15114–15119.
37. Kishi, N., and Macklis, J.D. (2004). MECP2 is progressively expressed in post-migratory neurons and is involved in neuronal maturation rather than cell fate decisions. Mol. Cell. Neurosci. *27*, 306–321.
38. Kishi, N., and Macklis, J.D. (2010). MeCP2 functions largely cell-autonomously, but also non-cell-autonomously, in neuronal maturation and dendritic arborization of cortical pyramidal neurons. Exp. Neurol. *222*, 51–58.
39. Kokura, K., Kaul, S.C., Wadhwa, R., Nomura, T., Khan, M.M., Shinagawa, T., Yasukawa, T., Colmenares, C., and Ishii, S. (2001). The Ski Protein Family Is Required for MeCP2-mediated Transcriptional Repression. J. Biol. Chem. *276*, 34115–34121.
40. Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science (80-.). *324*, 929–930.
41. Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., K. Beattie, B., Emili, A., Boone, C., et al. (2003). Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II. Mol. Cell. Biol. *23*, 4207–4218.
42. Lagger, S., Connelly, J.C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B.H., Yu, M., He, C., Sanguinetti, G., Sowers, L.C., et al. (2017). MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. PLOS Genet. *13*, e1006793.
43. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

44. 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.
45. Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science (80-.). *357*, 600 LP-604.
46. Lyst, M.J., and Bird, A. (2015). Rett syndrome: a complex disorder with simple roots. Nat. Rev. Genet. *16*, 261–275.
47. Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., de Lima Alves, F., et al. (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nat Neurosci *16*, 898–902.
48. Malik, A.N., Vierbuchen, T., Hemberg, M., Rubin, A.A., Ling, E., Couch, C.H., Stroud, H., Spiegel, I., Farh, K.K.-H., Harmin, D.A., et al. (2014). Genome-wide identification and characterization of functional neuronal activity–dependent enhancers. Nat. Neurosci. *17*, 1330.
49. McClelland, M., and Ivarie, R. (1982). Asymmetrical distribution of CpG in an “average” mammalian gene. Nucleic Acids Res. *10*, 7865–7877.
50. 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.
51. Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L., and Bird, A.P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499–507.
52. Mellén, M., Ayata, P., and Heintz, N. (2017). 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. Proc. Natl. Acad. Sci. U. S. A. *114*, E7812–E7821.
53. 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., et al. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron *86*, 1369–1384.
54. Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat. Rev. Neurosci. *8*, 427.
55. Nguyen, S., Meletis, K., Fu, D., Jhaveri, S., and Jaenisch, R. (2007). Ablation of de novo DNA methyltransferase Dnmt3a in the nervous system leads to neuromuscular defects and shortened lifespan. Dev Dyn *236*, 1663–1676.
56. Nott, A., Cheng, J., Gao, F., Lin, Y.-T., Gjoneska, E., Ko, T., Minhas, P., Zamudio, A.V., Meng, J., Zhang, F., et al. (2016). Histone deacetylase 3 associates with MeCP2 to regulate FOXO and social behavior. Nat Neurosci *19*, 1497–1505.
57. Pacheco, N.L., Heaven, M.R., Holt, L.M., Crossman, D.K., Boggio, K.J., Shaffer, S.A.,

Flint, D.L., and Olsen, M.L. (2017). RNA sequencing and proteomics approaches reveal novel deficits in the cortex of Mecp2-deficient mice, a model for Rett syndrome. Mol. Autism *8*, 56.

58. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.
59. Raman, A.T., Pohodich, A.E., Wan, Y.-W., Yalamanchili, H.K., Lowry, W.E., Zoghbi, H.Y., and Liu, Z. (2018). Apparent bias toward long gene misregulation in MeCP2 syndromes disappears after controlling for baseline variations. Nat. Commun. *9*, 3225.
60. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.
61. Renthal, W., Boxer, L.D., Hrvatin, S., Li, E., Silberfeld, A., Nagy, M.A., Griffith, E.C., Vierbuchen, T., and Greenberg, M.E. (2018). Characterization of human mosaic Rett syndrome brain tissue by single-nucleus RNA sequencing. Nat. Neurosci. *21*, 1670–1679.
62. Rube, H.T., Lee, W., Hejna, M., Chen, H., Yasui, D.H., Hess, J.F., LaSalle, J.M., Song, J.S., and Gong, Q. (2016). Sequence features accurately predict genome-wide MeCP2 binding in vivo. Nat. Commun. *7*, 11025.
63. Samaco, R.C., Mandel-Brehm, C., McGraw, C.M., Shaw, C.A., McGill, B.E., and Zoghbi, H.Y. (2012). Crh and Oprm1 mediate anxiety-related behavior and social approach in a mouse model of MECP2 duplication syndrome. Nat. Genet. *44*, 206–211.
64. Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C.T., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407.
65. Sceniak, M.P., Lang, M., Enomoto, A.C., James Howell, C., Hermes, D.J., and Katz, D.M. (2016). Mechanisms of Functional Hypoconnectivity in the Medial Prefrontal Cortex of Mecp2 Null Mice. Cereb. Cortex (New York, NY) *26*, 1938–1956.
66. Schaukowitch, K., Joo, J.-Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.-K. (2014). Enhancer RNA facilitates NELF release from immediate early genes. Mol. Cell *56*, 29–42.
67. Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321.
68. Schübeler, D., MacAlpine, D.M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D.E., O'Neill, L.P., Turner, B.M., Delrow, J., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. *18*, 1263–1271.
69. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E.,

Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

70. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116.
71. Shepherd, G.M.G., and Katz, D.M. (2011). Synaptic microcircuit dysfunction in genetic models of neurodevelopmental disorders: Focus on Mecp2 and Met. Curr. Opin. Neurobiol. *21*, 827–833.
72. Skene, P.J., Illingworth, R.S., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Andrews, R., and Bird, A.P. (2010). Neuronal MeCP2 Is Expressed at Near Histone-Octamer Levels and Globally Alters the Chromatin State. Mol. Cell *37*, 457–468.
73. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.
74. 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.
75. Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.
76. Stroud, H., Su, S.C., Hrvatin, S., Greben, A.W., Renthal, W., Boxer, L.D., Nagy, M.A., Hochbaum, D.R., Kinde, B., Gabel, H.W., et al. (2017). Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. Cell *171*, 1151–1164.e16.
77. Sugino, K., Hempel, C.M., Okaty, B.W., Arnson, H.A., Kato, S., Dani, V.S., and Nelson, S.B. (2014). Cell-Type-Specific Repression by Methyl-CpG-Binding Protein 2 Is Biased toward Long Genes. J. Neurosci. *34*, 12877–12883.
78. Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol. Cell *10*, 1453–1465.
79. Tudor, M., Akbarian, S., Chen, R.Z., and Jaenisch, R. (2002). Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. Proc. Natl. Acad. Sci. U. S. A. *99*, 15536–15541.
80. Weinreb, C., and Raphael, B.J. (2016). Identification of hierarchical chromatin domains. Bioinformatics *32*, 1601–1609.
81. 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.

82. Yamada, T., Yang, Y., Valnegri, P., Juric, I., Abnousi, A., Markwalter, K.H., Guthrie, A.N., Godec, A., Oldenborg, A., Hu, M., et al. (2019). Sensory experience remodels genome architecture in neural circuit to drive motor learning. Nature *569*, 708–713.
83. Zhao, Y.T., Goffin, D., Johnson, B.S., and Zhou, Z. (2013). Loss of MeCP2 function is associated with distinct gene expression changes in the striatum. Neurobiol Dis *59*, 257–266.
84. Ziats, M.N., Grosvenor, L.P., and Rennert, O.M. (2015). Functional genomics of human brain development and implications for autism spectrum disorders. Transl. Psychiatry *5*, e665.

2.8 Figures

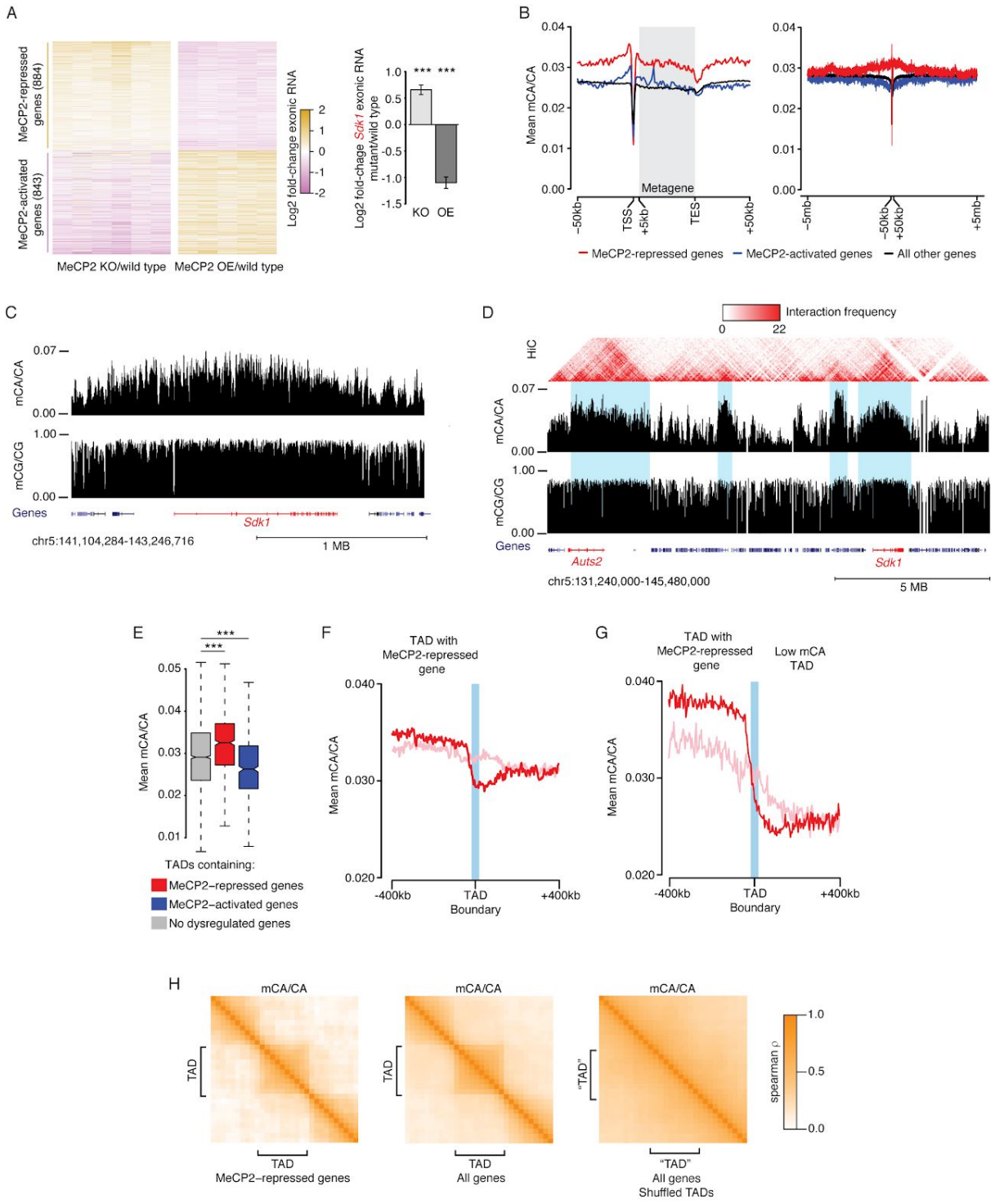
A
MeCP2-repressed genes (884)
MeCP2-activated genes (843)
MeCP2 KO/wild type
MeCP2 OE/wild type
Log2 fold-change exonic RNA
Log2 fold-chage Sdk1 exonic RNA mutant/wild type
KO OE
B
Mean mCA/CA
Metagene
TSS
TES
MeCP2-repressed genes
MeCP2-activated genes
All other genes
C
mCA/CA
mCG/CG
Genes
Sdk1
chr5:141,104,284-143,246,716
1 MB
D
Interaction frequency
HiC
Auts2
Sdk1
chr5:131,240,000-145,480,000
5 MB
E
TADs containing:
MeCP2–repressed genes
MeCP2–activated genes
No dysregulated genes
F
TAD with MeCP2-repressed gene
TAD Boundary
G
TAD with MeCP2-repressed gene
Low mCA TAD
TAD Boundary
H
mCA/CA
TAD
MeCP2–repressed genes
All genes
"TAD"
Shuffled TADs
spearman ρ

Figure 2.1. Non-CG methylation in cerebral cortex is associated with domains of chromatin folding.

A. Exonic RNA changes for all genes dysregulated in MeCP2 KO and OE (FDR <0.1) (left), and an example MeCP2-repressed gene, *Sdk1* (right). ***, B-H adjusted p<10^-^{8} Wald test.
B. Aggregate mCA/CA levels for MeCP2-regulated genes. Mean mCA/CA for 1kb bins shown at kilobase (left) and megabase (right) scale. "Metagene" is 50 equally-sized bins within gene bodies.
C. Genome browser view of mC at *Sdk1*.
D. Hi-C interactions and mC for a genomic region including two MeCP2-repressed genes, *Sdk1* and *Auts2* (red). TAD-like structures visible in Hi-C interactions (blue).
E. mCA/CA for TADs containing MeCP2-regulated genes (see methods). ***, p < 10^-^{8} Wilcoxon test.
F. Aggregate mCA/CA at boundaries of TADs containing MeCP2-repressed genes. Analysis of true TADs (red) or TADs shuffled around MeCP2-repressed genes (pink, see methods). mCA/CA drop-off at true boundaries is significantly different from shuffled boundaries. p<10^-^{3} (see Figure S2C, methods).
G. Aggregate mCA/CA as in F, but for TADs containing MeCP2-repressed genes with the top 33% most differential mCA/CA levels compared to the neighboring TAD (see methods).
H. Cross-correlation of mCA/CA levels for regions in and around TADs (see methods) containing MeCP2-repressed genes, all genes, and shuffled control TADs.

See also Figure 2.8, 2.9

Data from cerebral cortex of 7-10 week-old animals. n=6 each (MeCP2 KO, wild type) and n=5 each (MeCP2 OE, wild type) for RNA-seq, n=2 wild type for DNA methylation (Stroud et al., 2017). Hi-C data (Dixon et al., 2012).

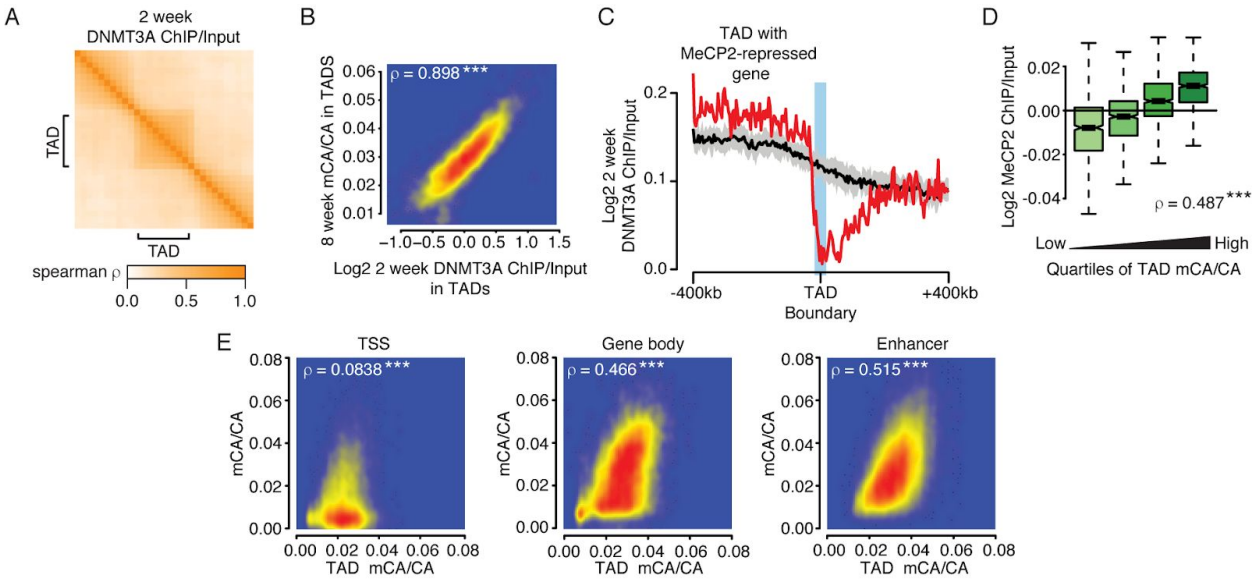

Figure 2.2. Domain-associated DNMT3A defines megabase- and kilobase-scale mCA levels.

A. Cross-correlations in and around TADs for DNMT3A ChIP at 2 weeks (see methods).
B. Comparison of DNMT3A ChIP/Input at 2 weeks and mCA/CA levels at 8 weeks for each TAD. ***, p<10^-^{8}.
C. Aggregate DNMT3A ChIP/Input at 2 weeks at boundaries of TADs containing MeCP2-repressed genes. Black line and ribbon indicate the mean and standard deviation of shuffled TAD boundaries (see methods).
D. MeCP2 ChIP/Input in TADs for quartiles of TAD mCA/CA. ***, p<10^-^{8}.
E. Comparison of TAD mCA/CA levels and mCA/CA levels at kilobase-scale genomic elements in each TAD at 8 weeks of age (see methods). ***, p<10^-^{8}.

See also Figure 2.10.

Data from cerebral cortex of 2 or 8 week-old animals. Per time point: n=2-3 for DNMT3A ChIP-seq, n=2 wild type for DNA methylation (Stroud et al., 2017).

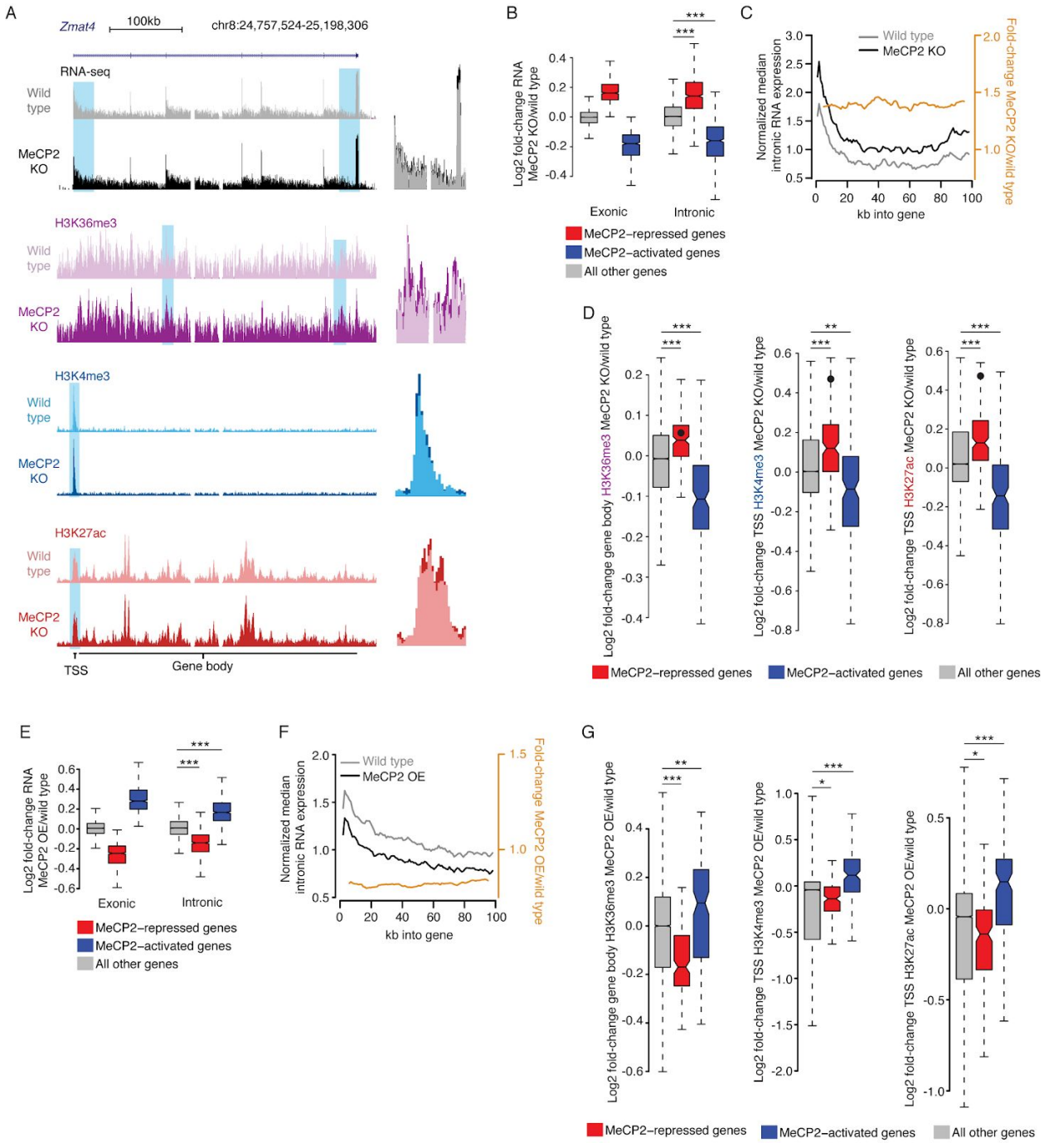

Figure 2.3. Disruption of MeCP2 leads to promoter-associated transcriptional dysregulation.

A. Left, genome browser view of nuclear total RNA-seq and ChIP-seq from MeCP2 KO and wild type at an MeCP2-repressed gene, *Zmat4*. Right, overlay of MeCP2 KO and wild type signal illustrating subtle increases for blue highlights at left.

B. Changes in RNA-seq signal in the MeCP2 KO for exons in whole cortex RNA (Exonic), or introns in nuclear RNA (Intronic). Genesets defined in combined analysis of exonic RNA (see Figure 1A). ***, $p<10^{-8}$ Wilcoxon test.
C. Profile (black, gray) and fold-change (orange) of intronic reads from RNA-seq for the first 100kb of upregulated genes in MeCP2 KO versus wild type (Figure S4B). Normalized median of 1kb bins is plotted for genes >100kb (see methods, and Figure S4G,H).
D. Fold-changes in ChIP signal in MeCP2 KO at MeCP2-regulated genes identified in intronic RNA analysis. Value for *Zmat4* is indicated by a point on each plot. **, $p<10^{-3}$; ***, $p<10^{-8}$ Wilcoxon test.
E. Fold-changes in gene expression as in panel B, but for MeCP2 OE. ***, $p<10^{-8}$ Wilcoxon test.
F. Profile of intron expression and fold-change as in panel C, but for MeCP2 OE. (see also Figure S4I).
G. Fold-changes in ChIP signal as in panel D but for MeCP2 OE. *, $p<0.05$; **, $p<10^{-3}$; ***, $p<10^{-8}$ Wilcoxon test.

See also Figure 2.11.

Data from cerebral cortex of 7-10 week-old animals. MeCP2 KO per genotype: n=6 for RNA-seq, n=3 for H3K36me3, n=4 for H3K4me3, n=5 for H3K27ac. MeCP2 OE per genotype: n=5 for RNA-seq, n=2 each for H3K36me3 and H3K4me3, n=3 for H3K27ac.

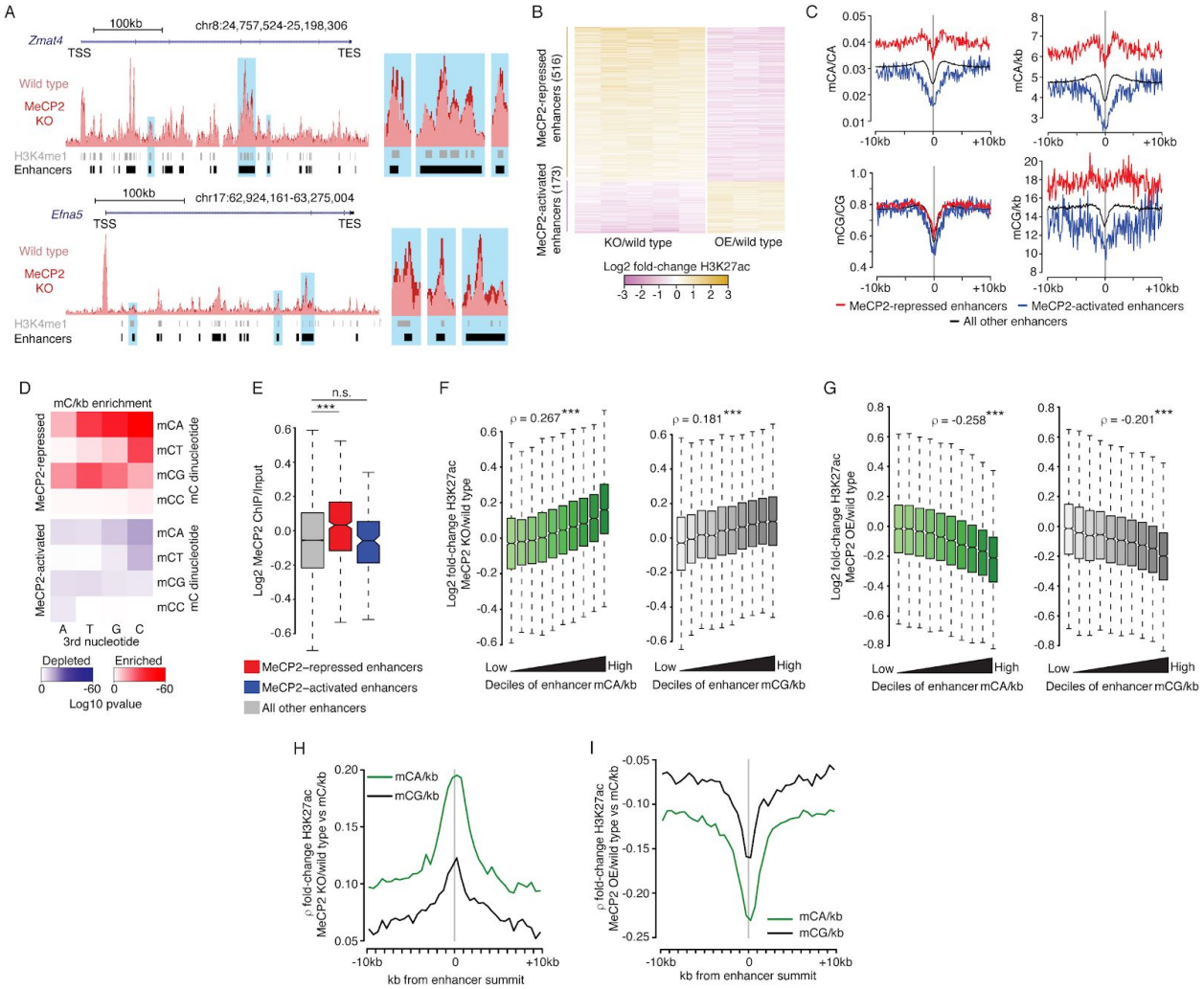

Figure 2.4. MeCP2 represses enhancers enriched for mCA and mCG binding sites.

A. Left, overlaid wild type and MeCP2 KO H2K27ac ChIP-seq signal at two MeCP2-repressed genes, *Zmat4* and *Efna5*. Right, close-up view of enhancers indicated in blue at left.

B. Fold-changes of H3K27ac ChIP in MeCP2 KO or MeCP2 OE for changed enhancers identified in combined analysis of H3K27ac ChIP-seq from MeCP2 KO and OE (FDR<0.1).

C. Aggregate mC at MeCP2-regulated enhancers, centered at the midpoint for each enhancer. Mean values plotted for 100 bp bins.

D. Enrichment significance (see methods) for mCNN trinucleotide density at MeCP2-repressed and MeCP2-activated enhancers.

E. MeCP2 ChIP/Input for MeCP2-regulated enhancers. ***, p<10^-^{8} Wilcoxon test.

F. Fold-changes of H3K27ac in MeCP2 KO across deciles of mCA/kb (left) and mCG/kb (right) for all enhancers. Spearman rho shown for mC and change in H3K27ac at enhancers. ***, p<10^-^{8}.

G. Fold-changes of H3K27ac, as in panel F, but for MeCP2 OE. ***, p<10^-^{8}.

H. Spearman correlations between H3K27ac fold-change at enhancers in the MeCP2 KO and mC/kb for 500bp bins across these enhancers. Plots centered at the summit of H3K27ac ChIP peaks (see methods).
I. Spearman correlations, as in panel H, but for H3K27ac fold-change in the MeCP2 OE.

See also Figure 2.12.

Data from cerebral cortex of 7-10 week-old animals. MeCP2 KO per genotype: n=5 for H3K27ac. MeCP2 OE per genotype: n=3 for H3K27ac. n=2 wild type for DNA methylation (Stroud et al., 2017) and MeCP2 ChIP-seq data (Kinde et al., 2016).

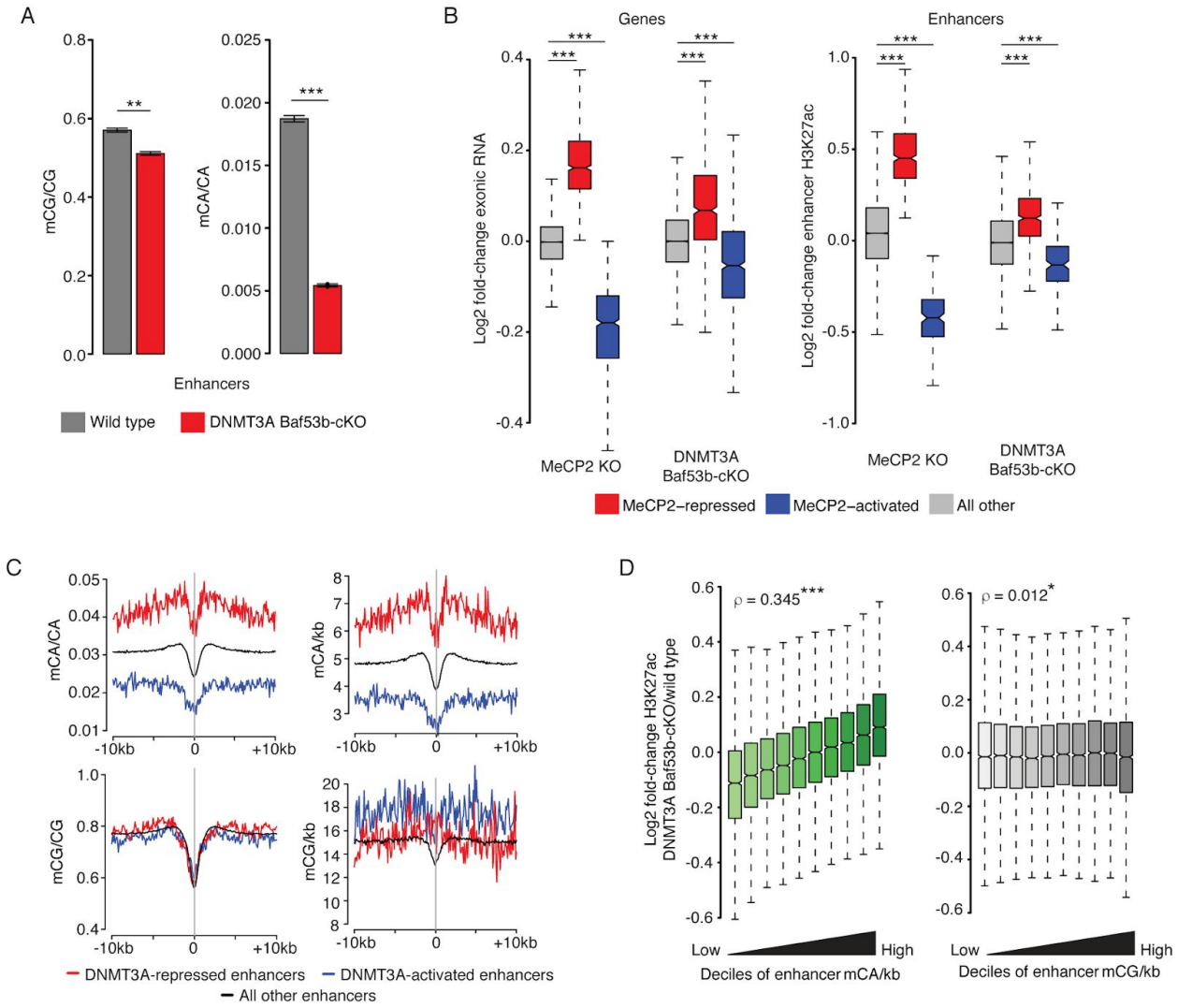

Figure 2.5. mCA-associated enhancer de-repression in DNMT3A Baf53b-cKO.

A. mCG/CG and mCA/CA levels at all enhancers for wild type and DNMT3A Baf53b-cKO. Data shown are mean +/- SEM. **, $p<10^{-3}$; ***, $p<10^{-8}$ two-tailed t-test.

B. Fold-changes for mRNA of MeCP2-regulated genes (left) and H3K27ac at MeCP2-regulated enhancers (right) in MeCP2 KO and DNMT3A Baf53b-cKO. ***, $p<10^{-8}$ Wilcoxon test.

C. Aggregate wild-type mC profiles for enhancers significantly dysregulated in the DNMT3A Baf53b-cKO (see Figure S6D).

D. Fold-changes of H3K27ac across deciles of mCA/kb (left) and mCG/kb (right) for all enhancers in the DNMT3A Baf53b-cKO. Spearman rho shown for mC and H3K27ac change at enhancers. *, $p<0.05$; ***, $p<10^{-8}$.

See also Figure 2.13.

Data from cerebral cortex of 7-10 week-old animals. Per genotype: n=6 for DNA methylation, n=4 for RNA-seq, n=6 for H3K27ac ChIP-seq.

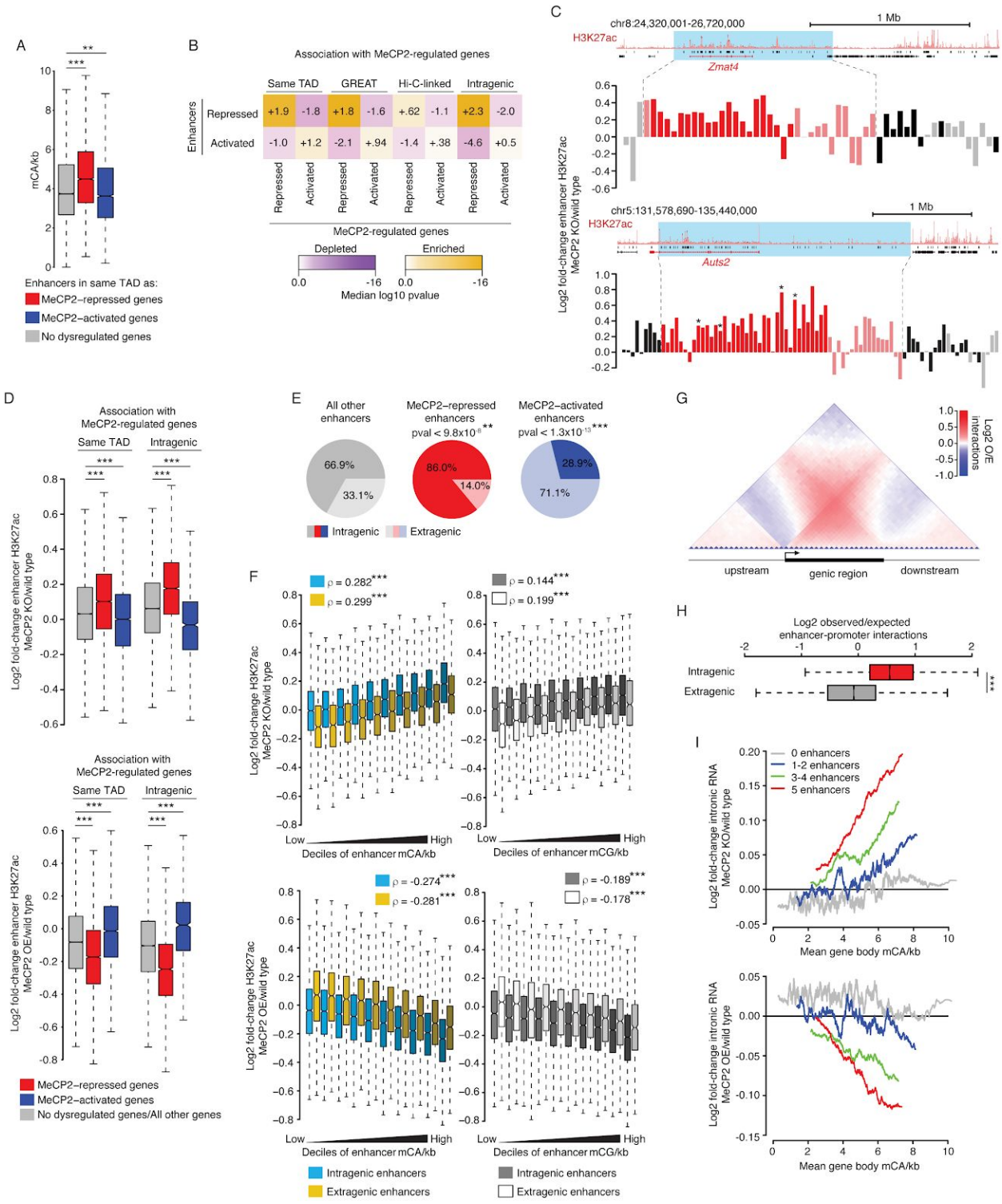

Figure 2.6. MeCP2-repressed enhancers are linked to MeCP2-repressed genes.

A. mCA/kb of enhancers in TADs containing MeCP2-regulated genes. **, $p<10^{-3}$; ***, $p<10^{-8}$ Wilcoxon test.
B. Associations between MeCP2-regulated enhancers and MeCP2-regulated genes. Comparisons are: enhancers in the same TAD as dysregulated genes ("Same TAD"), assigned to dysregulated genes by GREAT analysis (McLean et al., 2010) ("GREAT"), enriched for enhancer-promoter contacts by Hi-C ("Hi-C-linked"), and found in dysregulated genes ("Intragenic"). Median significance (color) and log2 enrichment (number) are shown for true enhancers compared to resampled control enhancers (see methods).
C. Log2 fold-changes of H3K27ac in MeCP2 KO for enhancers near two MeCP2-repressed genes. Blue, TADs overlapping MeCP2-repressed genes; red, enhancers in MeCP2-repressed gene; pink, other enhancers in the same TAD; gray and black, extragenic and intragenic enhancers in other TADs. *, significantly changed enhancers.
D. Fold-changes in H3K27ac ChIP in MeCP2 KO (top) and MeCP2 OE (bottom) for enhancers in TADs containing MeCP2-repressed, MeCP2-activated, or no dysregulated genes ("Same TAD"), or for enhancers in these genes ("Intragenic"). ***, $p<10^{-8}$ Wilcoxon test.
E. Genic distributions of MeCP2-regulated enhancers. **, $p<10^{-3}$; ***, $p<10^{-8}$ chi-squared test (see methods).
F. Fold-changes of H3K27ac ChIP in MeCP2 KO (top) or MeCP2 OE (bottom) for intragenic and extragenic enhancers across deciles of enhancer mCA/kb (left) and mCG/kb (right). Spearman rho for correlation of enhancer mC and change in H3K27ac. ***, $p<10^{-8}$.
G. Aggregate observed/expected Hi-C interaction frequencies for regions inside and outside of genes (see methods).
H. Observed/expected Hi-C interactions between enhancers and promoters for intragenic or distance-matched extragenic enhancers (see methods). ***, $p<10^{-8}$ Wilcoxon test.
I. Running-average plot of intronic RNA fold-change in the MeCP2 KO (top) and MeCP2 OE (bottom) versus gene body mCA/kb for genes containing various numbers of enhancers. Mean changes are plotted for genes sorted by gene body mCA/kb (bins of 201 genes with a 1-gene step).

See also Figure 2.12

Data from cerebral cortex of 7-10 week-old animals. MeCP2 KO per genotype: n=6 for RNA-seq, n=5 for H3K27ac. MeCP2 OE per genotype: n=3 for H3K27ac, n= 5 for RNA-seq. n=2 wild type for DNA methylation (Stroud et al., 2017). Hi-C data from E14.5 cortical neurons (Bonev et al., 2017).

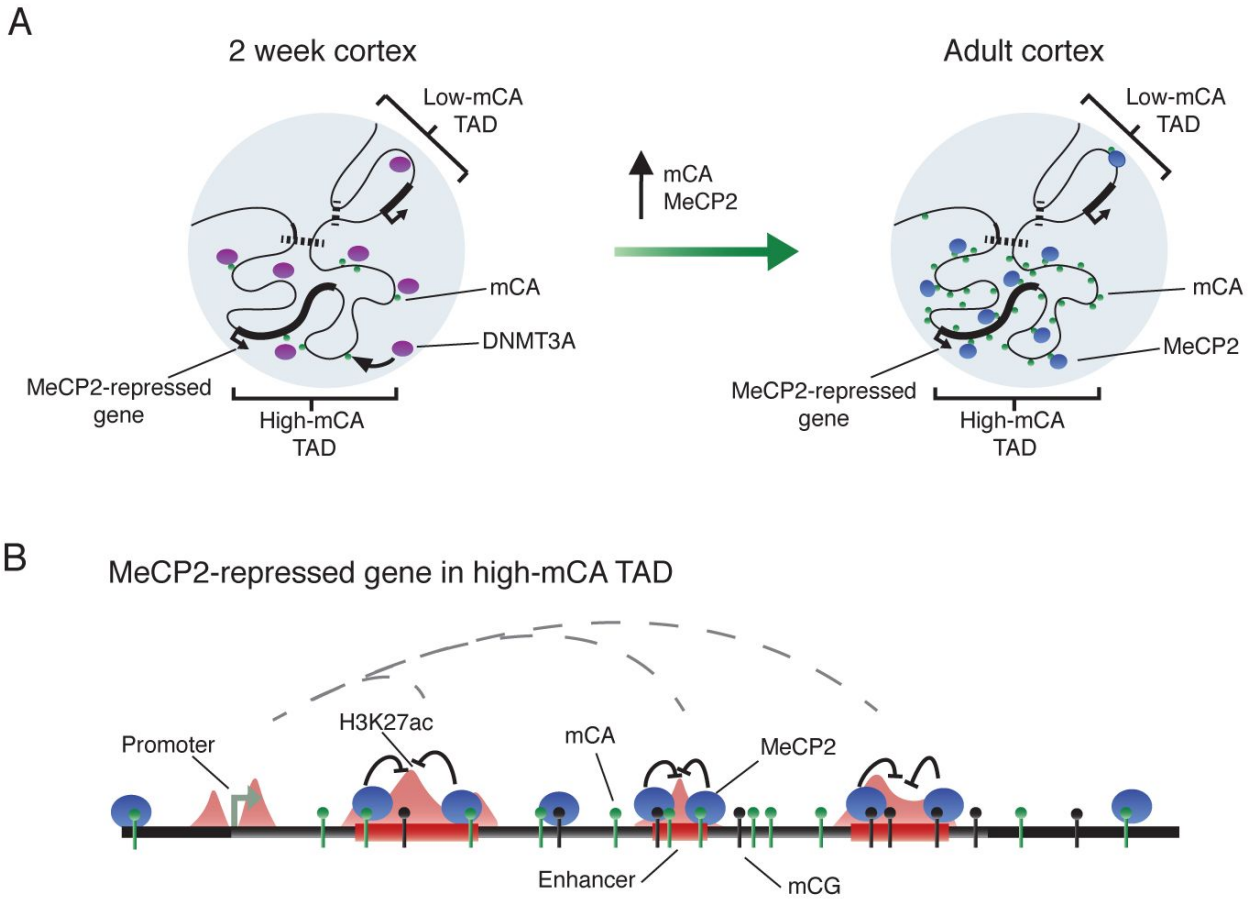

Figure 2.7. A model of TAD-associated mCA and enhancer repression by MeCP2 in neurons.

A. DNMT3A activity during early postnatal development establishes high or low mCA set-points within TADs.

B. In mature neurons, MeCP2 reads-out this mCA and mCG, repressing enhancers. MeCP2 most strongly affects intragenic enhancers, resulting in repression of genes found within high-mCA TADs that contain multiple enhancers.

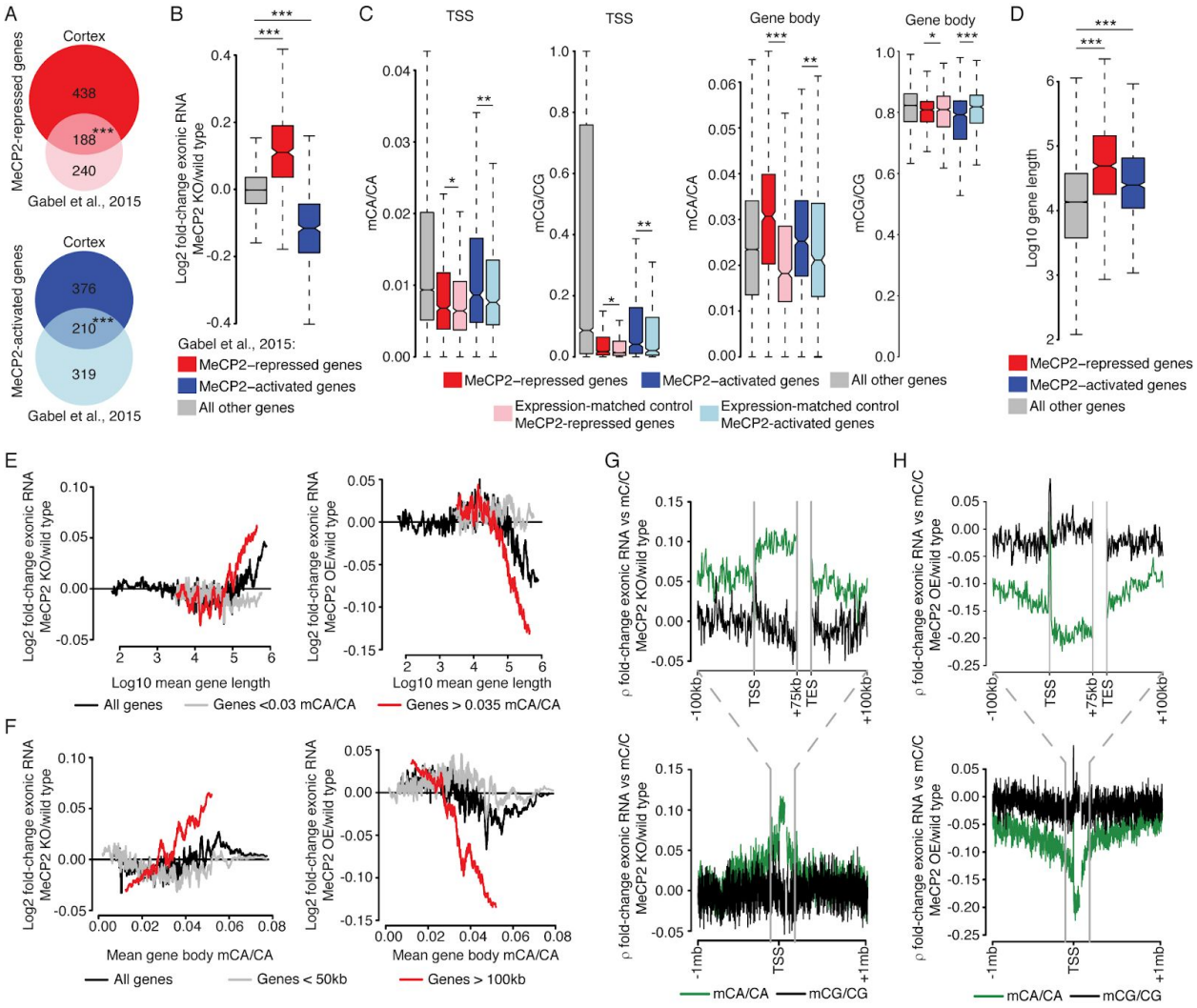

Figure 2.8. Identification of MeCP2-regulated genes in the cerebral cortex by combined RNA-seq analysis of MeCP2 KO and MeCP2 OE. Related to Figure 2.1.

A. Venn diagram of the overlap between genes identified as MeCP2-repressed and MeCP2-activated in combined analysis of MeCP2 KO and OE versus wild type cerebral cortex and genes previously identified as consistently dysregulated across multiple datasets and brain regions (Gabel et al., 2015). ***, p < 10^-8 hypergeometric test. Analysis was performed using coding genes found in annotation sets of both studies.

B. Boxplot of fold-changes in exonic RNA in the cortex of MeCP2 KO and wild type mice, for genes previously identified as consistently dysregulated across multiple datasets and brain regions (Gabel et al., 2015). ***, p < 10^-8 Wilcoxon rank-sum test.

C. Boxplots of mCA/CA and mCG/CG levels at the TSS (left) and gene bodies (right) of MeCP2-repressed, MeCP2-activat- ed, and all other genes. Light colored boxplots are plots generated for a control set of genes matched for the distribution of gene expression

for each dysregulated gene set (see methods). * p < 0.05 ; **, p <10^-3; ***, p < 10^-8 Medians of Wilcox- on rank-sum tests on 100 gene-set resamplings.

D. Boxplot of gene lengths for MeCP2-repressed, MeCP2-activated, and all other genes identified in combined analysis of total RNA-seq from MeCP2 KO and OE cerebral cortex. ***, p<10^-8 Wilcoxon rank-sum test.
E. Running average plots of exonic RNA fold-changes versus gene length, for MeCP2 KO (left) and OE (right) versus wild type. Mean fold-changes are plotted for bins of 201 genes sorted by gene length with a 1-gene step (see methods).
F. Running average plots of exonic RNA fold-change vs gene body mCA/CA for the MeCP2 KO (left) and OE (right) versus wild type. mCA/CA levels are calculated for the gene body defined as +3kb from the TSS to the TES. Mean fold-changes are plotted for bins of 201 genes sorted by mean mCA/CA per gene with a 1-gene step (see methods).
G. Plot of genome-wide correlations between methylation levels for 1kb regions in and around genes and fold-changes in exonic RNA expression in the MeCP2 KO versus wild type for each gene.
H. Plot of genome-wide correlations between methylation levels for 1kb regions in and around genes and fold-changes in exonic RNA expression in the MeCP2 OE versus wild type for each gene.

Data from cerebral cortex of 7-10 week old animals. n=6 per genotype for RNA-seq (MeCP2 KO, wild type), n=5 per genotype for RNA-seq (MeCP2 OE, wild type), n=2 wild type for DNA methylation (Stroud et al., 2017). In G and H analysis was carried out for genes over 75kb to visualize correlation signal within gene bodies.

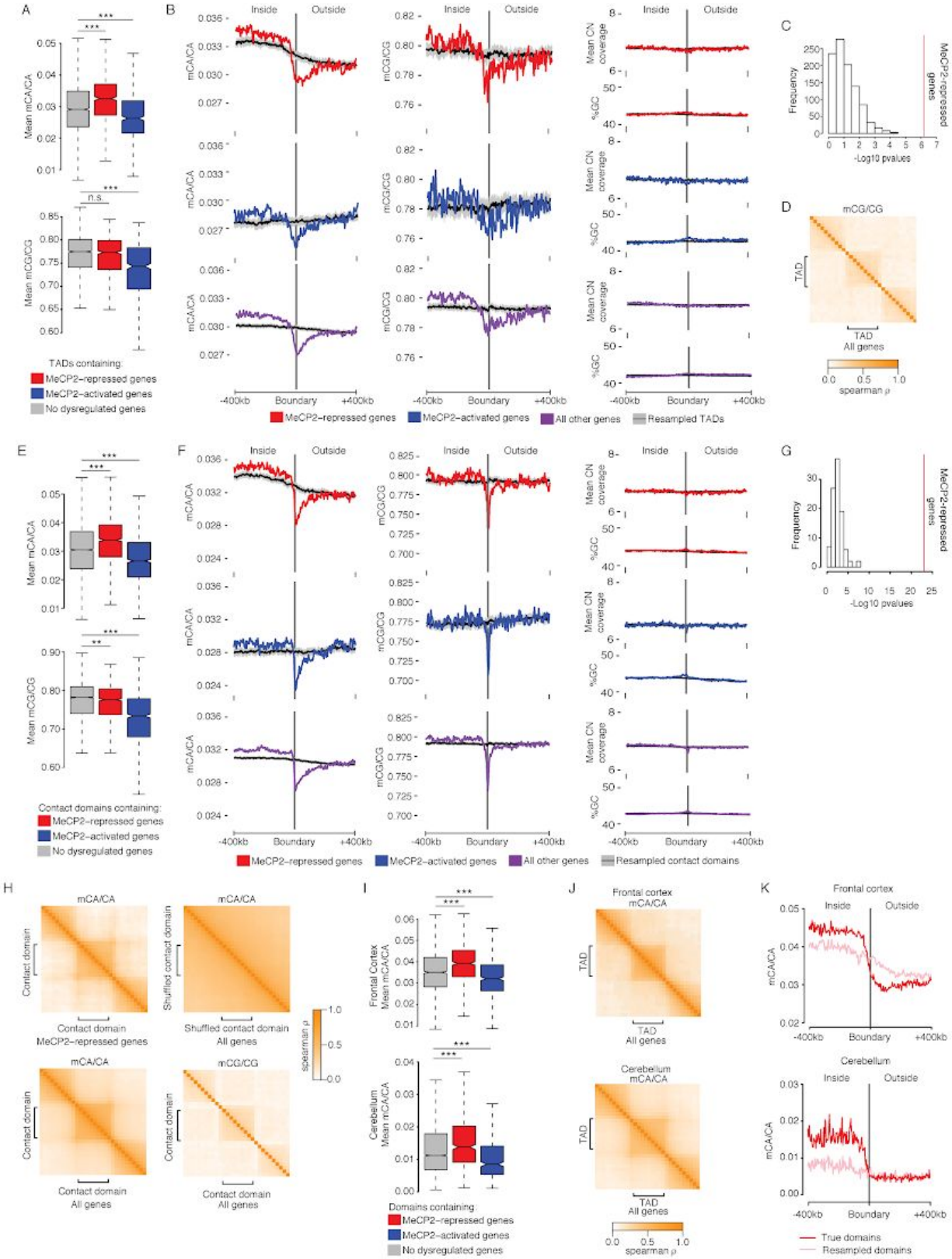
A
B
C
D
E
F
G
H
I
J
K
Mean mCA/CA
Mean mCG/CG
TADs containing:
MeCP2-repressed genes
MeCP2-activated genes
No dysregulated genes
n.s.
Inside
Outside
mCA/CA
mCG/CG
Mean CN coverage
%GC
-400kb
Boundary
+400kb
All other genes
Resampled TADs
Frequency
-Log10 pvalues
MeCP2-repressed genes
mCG/CG
TAD
All genes
spearman ρ
Contact domains containing:
Resampled contact domains
mCA/CA
Contact domain
MeCP2-repressed genes
Shuffled contact domain
All genes
mCG/CG
Frontal Cortex
Mean mCA/CA
Cerebellum
Mean mCA/CA
Domains containing:
Frontal cortex
Cerebellum
True domains
Resampled domains

Figure 2.9. Chromatin topology is associated with non-CG DNA methylation in the cerebral cortex. Related to Figure 2.1.

A. Boxplots of mCA/CA and mCG/CG within TADs that contain MeCP2-repressed genes, MeCP2-activated genes, or no dysregulated genes. ***, p < 10^-8 Wilcoxon rank-sum test.
B. Aggregate plots of mCA/CA (left), mCG/CG (center), and BS-sequencing coverage rates and GC composition (right) phased on boundaries of TADs defined in Hi-C data from the cerebral cortex at eight weeks of age. TADs containing MeCP2-repressed (red), MeCP2-activated (blue), or all other genes (purple) are shown. “Inside” indicates TADs containing genes of interest. Black line and ribbon for each plot indicates the mean and standard deviation of 20 sets of resampled boundaries generated by shuffling TAD locations in the genome (see methods).
C. A histogram of -log10 p-values for paired t-tests, comparing the difference in DNA methylation on either side of the TAD boundaries or 1000 resampled TAD boundaries as in B. Histogram shows the distribution of p-values for resampled TADs, red line indicates p-value for true TADs.
D. Cross correlation analysis of mCG/CG signal within and across TAD boundaries for all genes. mCG/CG values were calculated for 10 intra-domain regions and 10 equally-sized regions up and downstream of each TAD. Correlation between these regions across all TADs is shown (see methods).
E. Boxplots of mCA/CA and mCG/CG within contact domains that contain MeCP2-repressed genes, MeCP2-activated genes, or no dysregulated genes. Contact domains were defined by analysis of Hi-C data generated from cerebral cortex neurons isolated from fetal brain (Bonev et al., 2017). **, p < 10^-3; ***, p < 10^-8 Wilcoxon rank-sum test.
F. Aggregate plots of mCA/CA (left), mCG/CG (center), and BS-sequencing coverage rates and GC composition (right) phased on boundaries of contact domains defined in Hi-C data from fetal cortical neurons. Presented as in panel B.
G. A histogram of -log10 p-values for paired t-tests, comparing the difference in DNA methylation on either side of contact domain boundaries or 100 resampled contact domain boundaries as in panel F. Histogram shows the distribution of p-values for resampled contact domains, red line indicates p-value for true contact domains.
H. Cross correlation analysis of mCA/CA and mCG/CG signal within and outside of contact domains as in panel D. Resampling was performed by shuffling contact domain-sized regions around the genome and repeating the analysis of all genes (see methods).
I. Boxplots of mCA/CA within TADs defined in cerebral cortex (top) and cerebellum (bottom) that contain MeCP2-re- pressed genes, MeCP2-activated genes, or no dysregulated genes. ***, p < 10^-8 Wilcoxon rank-sum test. Cerebral cortex TADs were

compared to genes from this study. Cerebellum TADs were compared to genes previously identified as misregulated in the cerebellum and multiple other brain regions (Gabel et al., 2015).

J. Cross correlation analysis of mCA/CA signal within and outside of TADs defined in cerebral cortex (top) and cerebellum (bottom), as performed in panel H.
K. Aggregate plots of mCA/CA from frontal cortex (top) and granule neurons (bottom) phased on boundaries of MeCP2 repressed TADs defined in cerebral cortex (top) and cerebellum (bottom). Analysis performed on TADs with highly differing mCA levels as in Figure 2.1G.

A-D analysis of Hi-C interaction data (Dixon et al., 2012), and DNA methylation, n=2 (Stroud et al., 2017) from the cerebral cortex at 8 weeks of age. E-H analysis of Hi-C interaction data from neurons isolated from E14.5 cortex (Bonev et al., 2017), and DNA methylation, n=2 (Stroud et al., 2017) from the cerebral cortex at 8 weeks of age. I-K analysis of bisulfite data from 6-week frontal cortex and granule neurons isolated from 7-12 week old cerebellum (Lister et al., 2013; Mellén et al., 2017)(Mellén et al., 2017). Hi-C data from 6-8 week old Cerebellum (Yamada et al., 2019). Note that some panels from Figure 2.1 are repeated here to allow for comparisons.

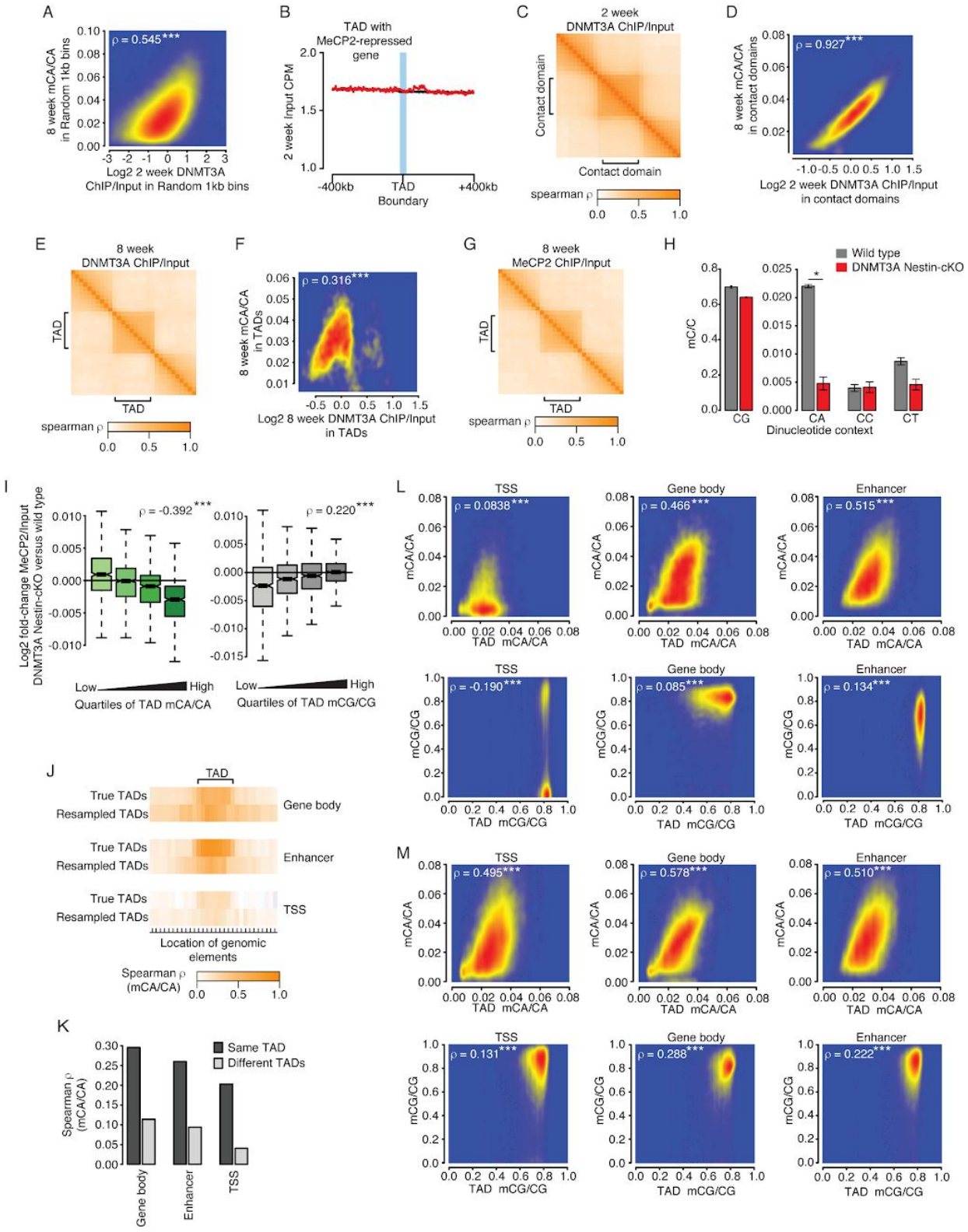

Figure 2.10. Binding of DNMT3A and recruitment of MeCP2 are shaped by topologically-associating domains. Related to Figure 2.2.

A. Density scatter plot of DNMT3A ChIP/Input signal at 2 weeks of age and mCA/CA levels at 8 weeks of age for random 1kb regions of the genome. Spearman rho shown for the correlation between DNMT3A signal and mCA/CA level. ***, p < 10^-8.
B. Aggregate plot of Input CPM at 2 weeks phased on boundaries of TADs that contain MeCP2-repressed genes (TADS from cerebral cortex at eight weeks of age). Black line and ribbon for each plot indicates the mean and standard deviation of 20 resampled boundaries generated by shuffling contact domain locations in the genome (see methods).
C. Cross correlation analysis of DNMT3A ChIP-seq signal at 2 weeks of age for contact domains defined in fetal cortical neurons. DNMT3A ChIP/Input values were calculated for 10 intra-domain regions and 10 equally-sized regions up and downstream of each domain. Correlation between these regions across all domains is shown (see methods).
D. Density scatter plot of DNMT3A ChIP/Input signal at 2 weeks of age and mCA/CA levels at 8 weeks of age for contact domains defined in fetal cortical neurons. Spearman rho shown for the correlation between DNMT3A signal and mCA/CA level. ***, p < 10^-8.
E. Cross correlation analysis (performed as in panel C) for DNMT3A ChIP-seq signal at 8 weeks of age in TADs defined in the cortex at 8 weeks of age.
F. Density scatter plot of DNMT3A ChIP/Input signal at 8 weeks of age and mCA/CA levels at 8 weeks of age for TADs defined in the cortex at 8 weeks of age. Spearman rho shown for the correlation between DNMT3A signal and mCA/CA level. ***, p < 10^-8.
G. Cross correlation analysis (performed as in panel C) for MeCP2 ChIP/Input signal at 8 weeks of age in TADs defined in the cortex at 8 weeks of age.
H. Barplots of genome-wide mC/C levels in a DNMT3A Nestin-cKO and control cerebral cortex at 8 weeks of age. Data shown as means +\- SEM. *, p < 0.05 two-tailed t-test.
I. Boxplot of fold-changes in MeCP2 ChIP-seq signal within TADs upon ablation of mCA in the DNMT3A Nestin-cKO. Change in signal is shown for TADs separated by quartiles of mCA/CA (left) and mCG/CG (right) under wild-type conditions. Spearman rho shown for the correlation between DNMT3A signal and mCA/CA level. *** p < 10^-8.
J. Heatmap of correlation between mCA/CA levels of gene body (top), enhancer (middle), and TSS (bottom) regions found inside and outside of TADs and the average mCA/CA level for each TAD. Distinct drop off in correlation coefficient for these elements when they are found outside the TAD boundary illustrates influence of intra-TAD mCA/CA consistency. Similar correlation level and drop-off in signal is not seen for resampled TADs, placed in randomized positions in the genome (see methods).
K. Barplots of correlations of mCA/CA levels for gene bodies (left), enhancers (middle), and TSSs (right), located in either the same TAD or separate TADs. To facilitate comparison, distances between pairs of elements in different TADs was matched to distances between pairs of elements in the same TAD (see methods).

L. Density scatter plots of TAD methylation levels and methylation at kilobase-scale genomic elements in the cerebral cortex at 8 weeks of age. Values for every TSS, gene body, or enhancer in the genome are plotted against the TAD in which the element resides. As in Figure 2E, data is subtracted from TADs such that element methylation does not contribute to plotted TAD methylation. Spearman rho shown for the methylation levels of each genomic element versus the methylation levels of the TAD that the element is in. ***, $p < 10^{-8}$.

M. Density scatter plots of TAD methylation levels and methylation at resampled random locations for regions size-matched to kilobase-scale genomic elements (as in panel L). Values for every resampled TSS-, gene body-, or enhancer-sized region in the genome are plotted against the TAD in which the element resides. Spearman rho shown for the methylation levels of each randomized genomic element versus the methylation levels of the TAD that the element is in. ***, $p < 10^{-8}$.

A,C,D Analysis of Hi-C interaction data from neurons isolated from E14.5 cortex (Bonev et al., 2017), DNA methylation, n=2, from the cerebral cortex at 8 weeks of age, and DNMT3A ChIP-seq data, n=3, from the cerebral cortex at 2 weeks of age (Stroud et al., 2017). B,E-G,I-M Analysis of Hi-C interaction data (Dixon et al., 2012), DNA methylation (n=2 wild type), DNMT3A ChIP-seq data, (n=2 wild type), (Stroud et al., 2017), and MeCP2 ChIP-seq data (n=2 per genotype) (Kinde et al., 2016) from the cerebral cortex at 8 weeks of age. H n=2 per genotype for DNA methylation DNMT3A Nestin-cKO versus wild-type, 8 weeks of age. Note that some panels from Figure 2.2 are repeated to allow for comparisons.

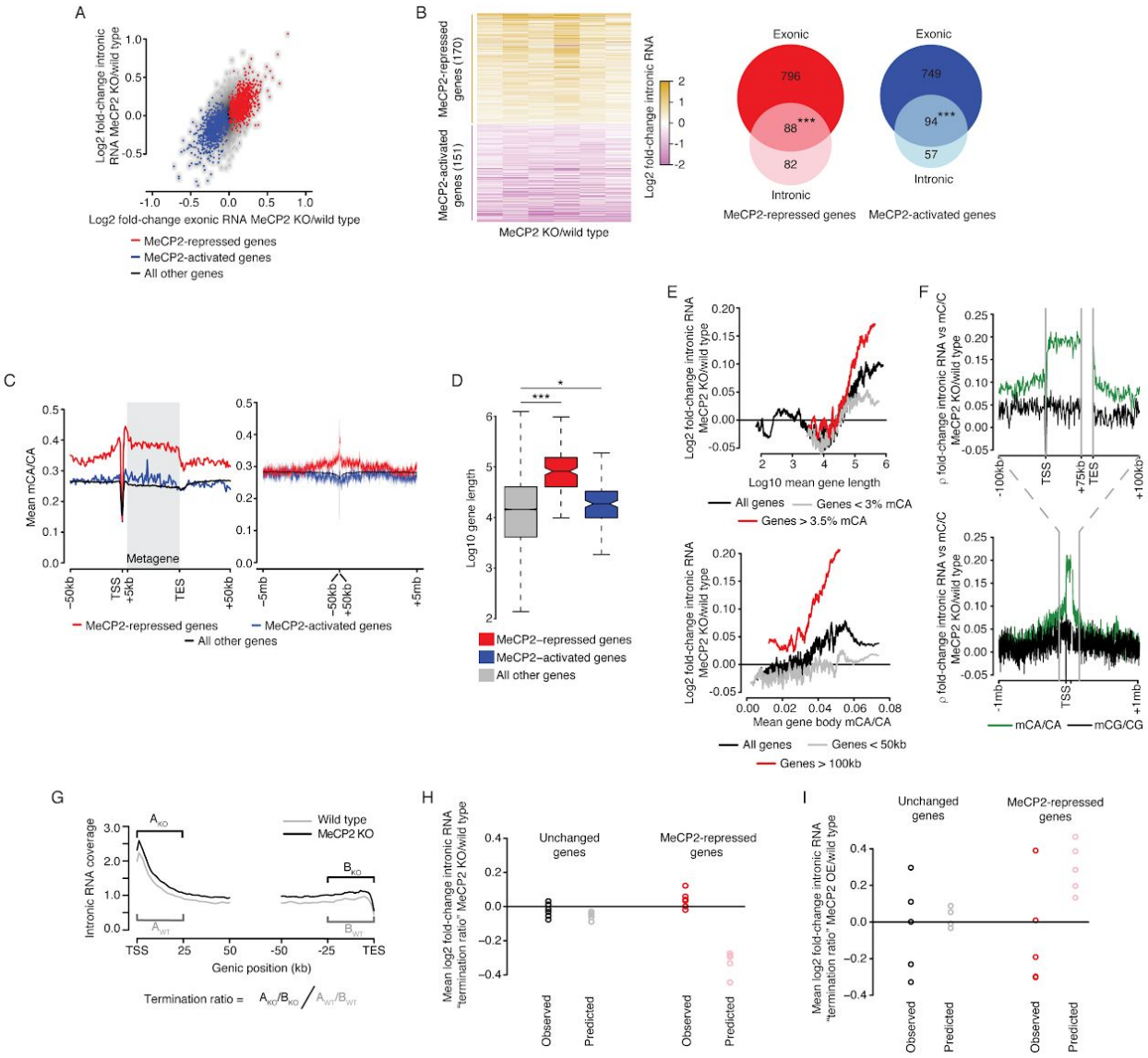

Figure 2.11. Changes in intronic RNA are consistent with promoter-associated transcriptional upregulation of long, highly methylated, MeCP2-repressed genes in the MeCP2 KO. Related to Figure 2.3.

A. Scatterplot of the log2 fold-changes in the MeCP2 KO versus wild type for exonic RNA measured by total RNA sequenc- ing of cerebral cortex tissue compared to intronic RNA measured by total RNA sequencing of isolated nuclei from this tissue. Genes identified as MeCP2-repressed and MeCP2-activated in combined analysis of exonic RNA in MeCP2 KO and MeCP2 OE mice (Figure 1; S1) are indicated by red and blue dots respectively, all other genes indicated as gray points.

B. Left, heatmap of changes in intronic RNA for genes detected as significantly dysregulated (FDR <0.1) in nuclear RNA-seq from MeCP2 KO cerebral cortex (grey indicates, no data for replicate). Right, venn diagram of the overlap between genes called

as MeCP2-repressed or MeCP2-activated in analysis of nuclear intronic RNA expression in the MeCP2 KO (pale colors) with genes called as dysregulated in combined analysis of exonic RNA in the MeCP2 KO and MeCP2 OE mice (dark colors) (Figure 1; S1). ***, p < 10^-8 hypergeometric test.

C. Aggregate plot of cerebral cortex mCA/CA levels for MeCP2-repressed, MeCP2-activated, and all other genes defined by analysis of intronic RNA in the MeCP2 KO (see panel B). Mean mCA/CA for 1kb bins in the TSS and regions surrounding genes is shown. For "Metagene" region, mean mCA/CA was calculated for 50 equal-sized bins within the body of each gene.
D. Boxplot of gene lengths for MeCP2-repressed or MeCP2-activated, and all other genes called by analysis of nuclear intronic RNA in the MeCP2 KO and wild type mice. *, p < 0.05; ***, p < 10^-8 Wilcoxon rank-sum test.
E. Running-average plot of fold-changes in intronic RNA expression in the MeCP2 KO and wild type versus mean gene length (top) or mean gene body mCA/CA (bottom). Mean fold-changes are plotted for bins of 201 genes sorted by length of mCA/CA per gene with a 1-gene step (see methods).
F. Plot of spearman correlations between fold-changes in intronic RNA expression in the MeCP2 KO and wild type versus mCA/CA or mCG/CG for regions in and around genes. Analysis performed for 1kb bins across 200kb (top) and 2 Mb (bottom) regions.
G. Scheme to assess changes in pre-mature transcription termination in the MeCP2 KO versus wild type. Similar to analysis carried out by Boswell et al., 2017, a "termination ratio" for each gene above 50kb in each sample is defined as the ratio of read counts in the first 25kb of each gene to the read counts in the last 25 kb of that gene. The mean fold-change in the ratios between MeCP2 KO and wild type is calculated across all genes in paired replicates.
H. Dotplot showing the change in termination ratio between the nuclear intronic RNA-seq data in the MeCP2 KO and wild type for genes that are not significantly changed and MeCP2-repressed genes. A prediction for the change in this ratio that would be expected if the effects on mRNA in the MeCP2 KO were due entirely to changes in transcription termina- tion rate was generated for comparison, "Predicted" (see methods).
I. Dotplot as in panel H, showing the change in termination ratio between the intronic RNA-seq in the MeCP2 OE and wild type for genes that are not significantly changed and MeCP2-repressed genes.

Data from cerebral cortex of 7-10 week old animals. For MeCP2 KO: n= 6 per genotype for RNA-seq; for MeCP2 OE: n=5 per genotype for RNA-seq; n=2 wild-type for DNA methylation (Stroud et al., 2017).

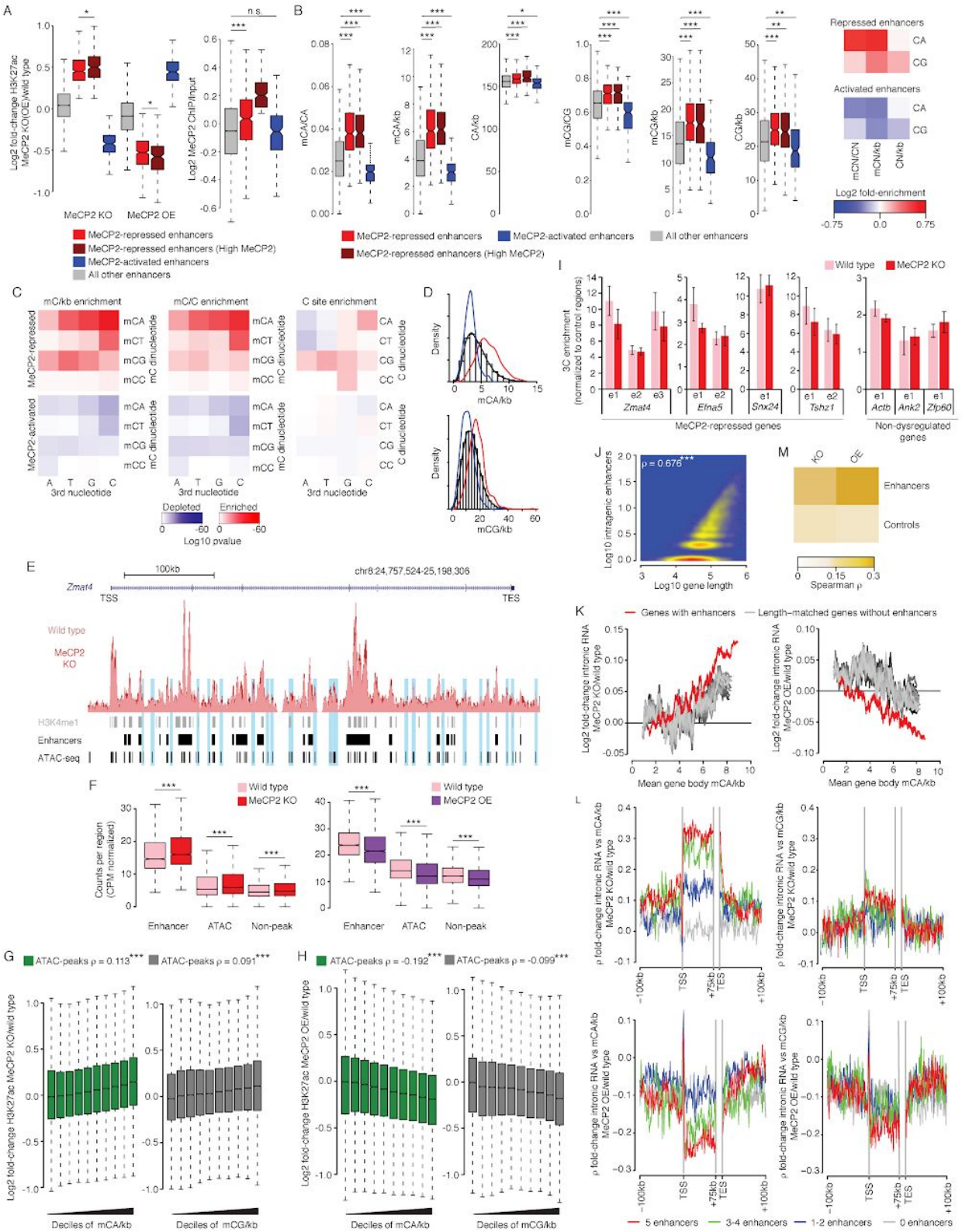
A
Log2 fold-change H3K27ac MeCP2 KO(OE)/wild type
MeCP2 KO
MeCP2 OE
Log2 MeCP2 ChIP/Input
n.s.
MeCP2-repressed enhancers
MeCP2-repressed enhancers (High MeCP2)
MeCP2-activated enhancers
All other enhancers
B
mCA/CA
mCA/kb
CA/kb
mCG/CG
mCG/kb
CG/kb
MeCP2-repressed enhancers
MeCP2-activated enhancers
All other enhancers
MeCP2-repressed enhancers (High MeCP2)
Repressed enhancers
Activated enhancers
CA
CG
mCN/CN
mCN/kb
CN/kb
Log2 fold-enrichment
C
mC/kb enrichment
mC/C enrichment
C site enrichment
MeCP2-repressed
MeCP2-activated
mCA
mCT
mCG
mCC
CA
CT
CG
CC
mC dinucleotide
C dinucleotide
3rd nucleotide
Depleted
Enriched
Log10 pvalue
D
Density
mCA/kb
mCG/kb
E
100kb
chr8:24,757,524-25,198,306
Zmat4
TSS
TES
Wild type
MeCP2 KO
H3K4me1
Enhancers
ATAC-seq
F
Wild type
MeCP2 KO
MeCP2 OE
Counts per region (CPM normalized)
Enhancer
ATAC
Non-peak
G
ATAC-peaks ρ = 0.113***
ATAC-peaks ρ = 0.091***
Log2 fold-change H3K27ac MeCP2 KO/wild type
Deciles of mCA/kb
Deciles of mCG/kb
H
ATAC-peaks ρ = -0.192***
ATAC-peaks ρ = -0.099***
Log2 fold-change H3K27ac MeCP2 OE/wild type
Deciles of mCA/kb
Deciles of mCG/kb
I
Wild type
MeCP2 KO
3C enrichment (normalized to control regions)
Zmat4
Efna5
Snx24
Tshz1
Actb
Ank2
Zfp60
MeCP2-repressed genes
Non-dysregulated genes
J
ρ = 0.676***
Log10 intragenic enhancers
Log10 gene length
M
KO
OE
Enhancers
Controls
Spearman ρ
K
Genes with enhancers
Length–matched genes without enhancers
Log2 fold-change intronic RNA MeCP2 KO/wild type
Log2 fold-change intronic RNA MeCP2 OE/wild type
Mean gene body mCA/kb
L
ρ fold-change intronic RNA vs mCA/kb MeCP2 KO/wild type
ρ fold-change intronic RNA vs mCG/kb MeCP2 KO/wild type
ρ fold-change intronic RNA vs mCA/kb MeCP2 OE/wild type
ρ fold-change intronic RNA vs mCG/kb MeCP2 OE/wild type
−100kb
TSS
+75kb
TES
+100kb
5 enhancers
3-4 enhancers
1-2 enhancers
0 enhancers

Figure 2.12. Analysis of enhancers dysregulated in MeCP2 mutants. Related to Figure 2.4 and Figure 2.6.

A. Boxplot of fold-changes in H3K27ac ChIP signal in MeCP2 KO and OE (left) and MeCP2 ChIP/Input signal (right) for enhancers identified with combined ChIP-seq analysis in KO and OE as MeCP2-repressed, MeCP2-repressed with high MeCP2/Input (see methods), MeCP2-activated, or all other enhancers. Fold-changes were calculated by edgeR analysis of H3K27ac ChIP-seq signal at enhancer regions (see methods). Significantly dysregulated enhancers were defined as FDR < 0.1. *, p < 0.05, ***, p < 10^-8 Wilcoxon rank-sum test.
B. Left, boxplots of mC/C, mC/kb, and C sites/kb at MeCP2-repressed, MeCP2-repressed with high MeCP2, MeCP2-activated, and all other enhancers, as displayed in panel A. Right, heatmap of enrichments for median levels of mC/C, mC/kb, and C sites/kb for MeCP2-repressed and MeCP2-activated enhancers compared to all other enhancers. *, p < 0.05, **, p < 10^-3; ***, p < 10^-8 Wilcoxon rank-sum test.
C. Heatmap of enrichment significance for number of mC sites (mC/kb, left), methylation level (mC/C, middle) and sequence occurrence (right) for trinucleotide sites in MeCP2-repressed and MeCP2-activated enhancers. Signifi- cance of enrichment or depletion was calculated by comparing occurrences of mC/kb, mC/C, or sequence alone for each trinucleotide in the 2kb region surrounding enhancers to resampled sets of enhancers that are not significantly changed and matched to changed enhancers for H3K27ac signal (see methods).
D. Histograms of mCA/kb and mCG/kb in enhancers (black). Blue and red lines indicate the distributions of methylation for MeCP2-activated and MeCP2-repressed enhancers respectively.
E. Genome browser view of an example MeCP2-repressed gene, Zmat4 (as in Figure 4A), showing overlaid wild-type and MeCP2 KO aggregate H3K27ac ChIP-seq signal, H3K4me1 ChIP-seq peaks, H3K27ac peaks called as enhancers, and peaks from a compendium of ATAC-seq peaks identified across 13 mouse tissues. Blue highlights ATAC-seq peaks that correspond to sub-thresholded enrichment of H3K27ac, not identified in peak calling analysis of H3K27ac ChIP-seq.
F. Boxplots of H3K27ac ChIP signal in MeCP2 KO versus wild type (left) and MeCP2 OE vs wild type (right) for enhancers defined in this study, a compendium of detectable ATAC-seq peaks, and non-peak regions that are size matched to ATAC-peaks but selected to not overlap enhancers of ATAC-seq peaks. Values for the highest mCA/CA decile are shown for each class, illustrating total levels of H3K27ac and mCA-dependent dysregulation ***, p < 10^-8 Wilcoxon rank-sum test.
G. Boxplots of fold-changes of H3K27ac ChIP signal in MeCP2 KO versus wild type across deciles of mCA/kb (left) and mCG/kb (right) for ATAC-peaks described in panel E.

Spearman rho shown for correlations of each methylation mark with change of H3K27ac at enhancers. ***, p < 10^-8 Wilcoxon rank-sum test.

H. Boxplots of fold-changes of H3K27ac ChIP signal in MeCP2 OE versus wild type across deciles of mCA/kb (left) and mCG/kb (right) for ATAC-peaks described in E. Spearman rho shown for correlations of each methylation mark with change of H3K27ac at enhancers. ***, p < 10^-8.

I. Quantitative 3C analysis of enhancers found in MeCP2 repressed genes, or control loci. Data shown as means +\- SEM. No significant factors were detected in a 2-way ANOVA, using enhancer locations and genotype.

J. Density scatter plot of log10 number of intragenic enhancers versus log10 gene length for all genes in the genome. Spearman rho shown for the number of intragenic enhancers and gene length for all genes with at least 1 enhancer. ***, p < 10^-8.

K. Running-average plot of fold-change in intronic RNA expression in the MeCP2 KO or MeCP2 OE versus mCA/kb for genes containing enhancers (red line) and length-matched sets of genes that do not contain enhancers (gray lines). Mean fold-changes are plotted for bins of 201 genes sorted by mCA/kb with a 1-gene step (see methods).

L. Local correlation analysis correlating mCA/kb (left) and mCG/kb (right) for 1 kb windows in and around genes to intronic RNA fold-changes for the gene in the MeCP2 KO (top) and OE (bottom). Genes at least 75kb in length are analyzed to allow visualization of correlations in gene bodies. Genic windows of methylation go from the TSS to 75kb downstream.

M. Heatmap of correlation between fold-change in H3K27ac at intragenic regions and gene expression fold-change for MeCP2 KO and MeCP2 OE mice. Values are calculated as partial correlation for enhancers or control non-enhancer regions of the same size, removing the signal from the other class of element (see methods). Larger correlation for enhancers over control regions illustrates the link between enhancer regulation and transcriptional control.

Data from cerebral cortex of 7-10 week old animals. For MeCP2 KO: n=5 per genotype for H3K27ac, n=6 per genotype for RNA-seq; for MeCP2 OE: n=3 per genotype for H3K27ac, n= 5 per genotype for RNA-seq; n=2 wild-type for DNA methylation (Stroud et al., 2017). ATAC-seq peaks (Cusanovich et al., 2018). Note that some panels in C are repeated from Figure 2.4 to allow for comparisons.

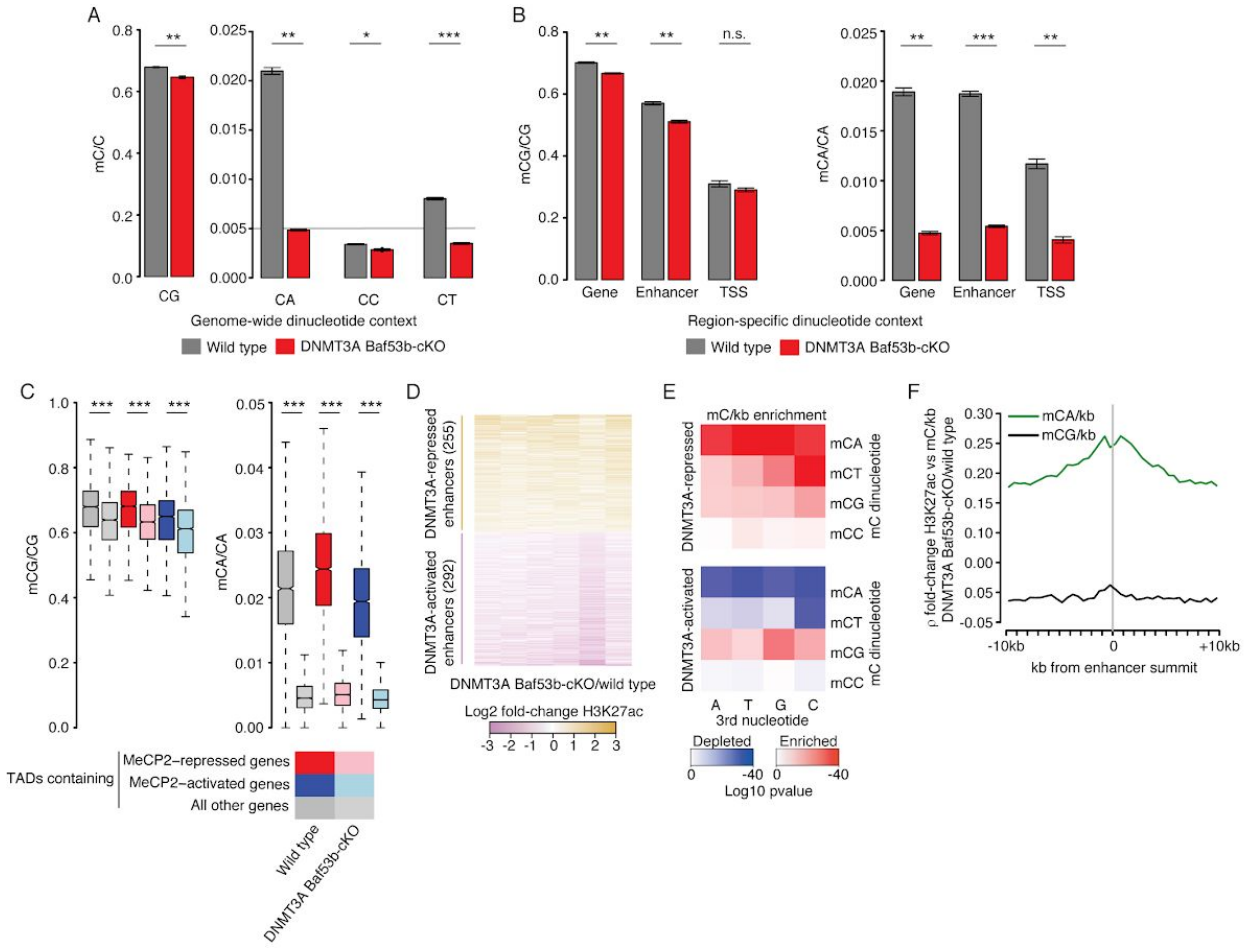

Figure 2.13. Transcriptomic and epigenomic analysis of DNMT3A Baf53b-cKO. Related to Figure 2.5.

A. Barplots of genome-wide methylation levels in the DNMT3A Baf53b-cKO at different CN dinucleotides. Data shown as means +\- SEM. *, p < 0.05, **, p < 10^-3; ***, p < 10^-8 two-tailed t-test.

B. Barplots of mCG (left) and mCA (right) levels in the DNMT3A Baf53b-cKO at genes, enhancers, and TSSs. Data shown as means +\- SEM. **, p < 10^-3; ***, p < 10^-8 two-tailed t-test.

C. Boxplots of mCG (left) and mCA (right) levels in the DNMT3A Baf53b-cKO within TADs that contain MeCP2-re- pressed genes, MeCP2-activated genes, or no dysregulated genes. ***, p < 10^-8 Wilcoxon rank-sum test.

D. Heatmap of H3K27ac fold-changes for enhancers detected as significantly dysregulated in DNMT3A Baf53b-cKO mice by edgeR analysis (see methods).

E. Heatmap of enrichment significance for the number of mC sites (mC/kb) within enhancers detected as significantly dysregulated in the DNMT3A Baf53b-cKO. Significance of enrichment or depletion was calculated by comparing occurrences of

mC/kb for each trinucleotide in the 2kb region surrounding enhancers to resampled sets of enhancers that are not significantly changed but matched to changed enhancers for H3K27ac signal (see methods).

F. Plot of spearman correlations between H3K27ac fold-change at enhancers in the DNMT3A Baf53b-cKO and mC/kb levels in the wild-type brain for 500bp bins in and around the enhancers. Plots are centered at the summit of enhancer H3K27ac ChIP peaks.

Data from cerebral cortex of 7-10 week old animals. n=6 per genotype for Bisulfite-seq analysis, n=6 per genotype for H3K27ac.

Chapter 3: DNMT3A haploinsufficiency results in behavioral deficits and global epigenomic dysregulation shared across neurodevelopment disorders

This chapter is adapted from a manuscript submitted to Cell Reports and published on Biorxiv:

Christian DL, Wu DY, Martin JR, Moore JR, Liu YR, Clemens AW, Nettles SA, Kirkland NM, Hill CA, Wozniak DF, Dougherty JD, Gabel HW. DNMT3A haploinsufficiency results in behavioral deficits and global epigenomic dysregulation shared across neurodevelopment disorders. bioRxiv

. 2020;77(2):279-293.e8. doi:10.1101/2020.07.10.195859

D.L.C and D.Y.W. are joint first authors, as each led critical components of the project and analysis. D.L.C., J.R.Ma., and Y.R.L. generated and analyzed in vitro biochemical data. Y.R.L. and S.A.N. generated primary neuronal culture samples. D.L.C. and J.R.Ma. generated skeletal samples. N.M.K. and C.A.H. carried out craniofacial analysis and D.L.C. carried out long bone analysis. D.L.C., J.R.Ma., T.P., D.F.W., and J.D.D. carried out behavioral tests and analysis. D.L.C., J.R.Ma., J.R.Mo., Y.R.L., and A.W.C. generated genomic data. D.Y.W. developed analysis algorithms and pipelines. D.L.C., D.Y.W., J.R.Mo, and A.W.C. completing genomic

analyses. H.W.G. conceived the project and H.W.G., D.L.C., and D.Y.W. designed the experiments. H.W.G., D.L.C., and D.Y.W. wrote the manuscript and all authors contributed to manuscript editing and revisions.

3.1 Abstract

Mutations in DNA methyltransferase 3A (DNMT3A) have been detected in autism and related disorders, but how these mutations disrupt nervous system function is unknown. Here we define the effects of DNMT3A mutations associated with neurodevelopmental disease. We show that diverse mutations affect different aspects of protein activity yet lead to shared deficiencies in neuronal DNA methylation. Heterozygous DNMT3A knockout mice mimicking DNMT3A disruption in disease display growth and behavioral alterations consistent with human phenotypes. Strikingly, in these mice we detect global disruption of neuron-enriched non-CG DNA methylation, a binding site for the Rett syndrome protein MeCP2. Loss of this methylation leads to enhancer and gene dysregulation that overlaps with models of Rett syndrome and autism. These findings define effects of DNMT3A haploinsufficiency in the brain and uncover disruption of the non-CG methylation pathway as a convergence point across neurodevelopmental disorders.

3.2 Introduction

Precise regulation of transcription through epigenetic mechanisms is critical for nervous system development (Cholewa-Waclaw et al., 2016). Exome sequencing studies have revealed mutations in genes encoding epigenetic modifiers of chromatin structure as major underlying

causes of neurodevelopmental diseases (NDD), including autism spectrum disorder (ASD) (McRae et al., 2017; Sanders et al., 2015; Satterstrom et al., 2019). A challenge emerging from these discoveries is to define the cellular functions of the disrupted proteins during normal development and to search for shared pathways between these proteins that can potentially be targeted for therapeutic development.

Gene regulation mediated by DNA methylation has emerged as an epigenetic mechanism that plays a critical role in nervous system function (Kinde et al., 2015). In addition to the classical methylation of cytosines found at CG dinucleotides (mCG), neurons contain uniquely high levels of methyl-cytosine (mC) in a non-CG context, with this mark occurring primarily at CA dinucleotides (mCA) (Guo et al., 2014; Lister et al., 2013; Xie et al., 2012). mCA is deposited *de novo* through the activation of the DNA methyltransferase 3A (DNMT3A) enzyme during the early postnatal period (1-6 weeks of age in mice). Levels of mCA increase specifically in neurons until the number of methylation sites in the non-CG context is nearly equivalent to the number of mCG sites (Guo et al., 2014; Lister et al., 2013; Xie et al., 2012). A critical function for mCA is to serve as a binding site for a neuron-enriched chromatin protein, Methyl-CpG binding Protein 2 (MeCP2) (Chen et al., 2015; Gabel et al., 2015; Guo et al., 2014). MeCP2 was initially defined by its high affinity for mCG, but biochemical and genomic studies indicate that it also tightly interacts with mCA to down-regulate transcription of genes with essential functions in the brain (Boxer et al., 2019; Gabel et al., 2015; Kinde et al., 2016; Lagger et al., 2017; Lyst and Bird, 2015). Loss of MeCP2 leads to the severe neurological disorder Rett syndrome, while duplication causes MeCP2-duplication syndrome, an ASD, suggesting that read-out of mCA is critical to nervous system function (Amir et al., 1999; Van Esch et al., 2005).

Human exome sequencing studies have recently identified *de novo* mutations in DNMT3A in individuals with ASD (Feliciano et al., 2019; Iossifov et al., 2014; Sanders et al., 2015; Satterstrom et al., 2019). Separate studies have also defined heterozygous disruption of DNMT3A as the underlying cause of Tatton-Brown Rahman syndrome (TBRS), a heterogeneous NDD characterized by intellectual disability, overgrowth, craniofacial abnormalities, anxiety, and high penetrance of ASD (Tatton-Brown et al., 2014, 2018). While a small portion of the mutations identified in affected individuals are truncations that are predicted to cause complete inactivation of the enzyme, a majority of disease-associated alleles are missense mutations, raising questions about whether loss-of-function effects are a primary mechanism of disruption in DNMT3A disorders (Tatton-Brown et al., 2014, 2018). In addition, the large percentage of missense mutations identified in affected individuals raises the need to establish deleterious effects of these mutations, helping to definitively associate the mutations with NDD and end the "diagnostic odyssey" for affected individuals and care-takers (Sawyer et al., 2016). Finally, while heterozygous loss of DNMT3A has been studied in the context of oncogenesis in the hematopoietic system (Cole et al., 2017), the effects of partial loss of DNMT3A on nervous system function *in vivo* have not been examined and the consequences of possible methylation changes on neuronal gene regulation and behavior are unknown.

Here we examine the molecular effects of neurodevelopmental disease-associated DNMT3A mutations and explore the consequences of heterozygous DNMT3A mutation on the neuronal epigenome. Our results indicate that missense mutations across canonical domains of DNMT3A disrupt different aspects of protein function, yet mutations in all domains reduce the enzyme's capacity to deposit neuronal mCA. We then use DNMT3A heterozygous deletion mice

as a model to assess the impact of shared DNMT3A loss-of-function effects that occur in affected individuals with missense and deletion mutations. We detect altered growth and behavior in these mice, supporting haploinsufficiency as a driver of pathology in DNMT3A disorders. Through integrated epigenomic analysis, we reveal disruption of mCA throughout the brain of DNMT3A mutant mice. We show that this loss of mCA leads to disruption of distal regulatory enhancer activity and changes in gene expression that overlap with models of MeCP2 disorders and other ASDs. These findings define the effects of NDD-associated DNMT3A mutations for the first time and implicate the disruption of mCA-mediated epigenomic regulation as a convergence site across clinically distinct NDDs.

3.3 Results

Functional analysis of disease-associated DNMT3A mutations

Multiple DNMT3A mutations have been identified in individuals with ASD and TBRS. However, the large number of missense mutations identified and the phenotypic heterogeneity of individuals with these mutations suggests that alterations of amino acids within different protein domains may have distinct consequences that dictate the nature and severity of disease. Furthermore, it is not clear if missense mutations identified in patient populations are truly disruptive, raising the need for direct assessment of molecular function to ascribe mutations as potentially causative. We therefore sought to assess the effects of disease-associated DNMT3A mutations on protein expression, cellular localization, and catalytic activity, looking for common effects that may be core to the development of NDD.

We engineered amino-acid alterations homologous to human disease mutations into a FLAG-tagged DNMT3A protein expression vector and assessed multiple mutations found within each functional domain of the protein (Figure 3.1A). These analyses included mutations in the chromatin interacting proline-tryptophan-tryptophan-proline (PWWP) domain, the auto-inhibitory Histone H3 lysine 4 interacting ATRX-DNMT3-DNMT3L (ADD) domain, and the well-defined methyltransferase catalytic domain (Gowher and Jeltsch, 2018). Transfection into heterologous cells facilitated rapid assessment of protein expression by western blot, cellular localization by immunocytochemistry, and catalytic activity by using an *in vitro* methyltransferase assay. Mutations in the PWWP domain resulted in a reduction in DNMT3A protein expression and loss of nuclear localization compared to wild-type controls (Figure 3.1B-D, Figure 3.8A-C). When expressed at equal levels to that of wild-type protein however, these mutations exhibited normal catalytic activity in the *in vitro* methyltransferase analysis (Figure 3.1E,F). In contrast, mutations found in the catalytic methyltransferase domain of DNMT3A showed wild-type expression and localization but displayed deficits in catalytic activity *in vitro* (Figure 3.1B-F, Figure 3.8A-C). Mutations in the ADD domain of DNMT3A displayed normal protein localization and expression levels and exhibited equal or higher methylation activity compared to wild-type protein *in vitro* (Figure 3.1B-F, Figure 3.8A-C).

To further evaluate the functional effects of disease-associated DNMT3A mutations in the context of endogenous chromatin, we tested the capacity of DNMT3A mutants to establish DNA methylation across the genome in mouse cortical neurons. For this analysis, we focused on the DNMT3A-dependent global build-up of mCA in postmitotic neurons (Gabel et al., 2015; Guo et al., 2014; Lister et al., 2013). Cultured neurons isolated from the cerebral cortex at

embryonic day 14.5 accumulate mCA *in vitro* and this build-up can be blocked by lentiviral-mediated delivery of Cre recombinase to DNMT3A$^{flx/flx}$ cells at 3 days *in vitro* (DIV) (Figure 3.2A, Figure 3.8D,E). We co-transduced Cre with wild-type or mutant DNMT3A lentivirus at equal levels (Figure 3.8F) to test the capacity of each protein to rescue deposition of DNA methylation. Testing of mutations across the major domains of DNMT3A detected deficits in mCA accumulation for all disease-associated mutations tested (Figure 3.2B). In this analysis, mutations in the ADD domain that exhibited robust catalytic activity *in vitro* displayed moderate-to-severe deficits in mCA deposition in neurons. The ADD domain has been implicated in both histone binding and auto-inhibition of the protein (Guo et al., 2015), thus results in this neuronal assay may indicate that loss of ADD function blocks the capacity of the enzyme to engage with chromatin and promote DNMT3A methylation activity in cells. Notably, disease mutations that may be associated with more severe patient phenotypes (intellectual disability) show a near-complete loss of mCA rescue, while mutations associated with ASD but not ID (such as V665L) have a more moderate depletion of mCA. This suggests a possible gradient of disease phenotypes that may be driven by differential mutation severity. Unlike mCA, global mCG levels show no significant differences in this *in vitro* system, likely due to the redundant function of the DNMT1 methyltransferase in maintaining methylation of CG sites (Figure 3.8G,H). Together our results indicate that although NDD-associated mutations in DNMT3A affect different protein domains and alter distinct aspects of protein function (e.g. localization, chromatin interaction, catalysis), these mutations share a common outcome of reduced enzymatic activity on neuronal DNA, with many mutations resulting in functionally null proteins.

In vivo growth and behavioral effects of heterozygous DNMT3A disruption

Our *in vitro* findings that multiple NDD-associated missense mutations in DNMT3A result in complete or near-complete loss of function, coupled with the identification of deletion and nonsense mutations in TBRS patients, suggest that 50% reduction of DNMT3A activity occurs in a substantial fraction of affected individuals. Thus, studying the effects of heterozygous inactivation of DNMT3A *in vivo* is an important first step in understanding the potential epigenomic and systems level consequences of DNMT3A disruption in NDD. Previous studies have demonstrated severe developmental deficits, disruptions in movement, and perinatal lethality associated with constitutive complete loss of DNMT3A (homozygous null mutation) (Okano et al., 1999) and with removal of DNMT3A from the brain (conditional DNMT3A deletion) (Nguyen et al., 2007). However, the relevance of heterozygous mutation of DNMT3A to neurodevelopmental disease has only recently been uncovered, and the growth and behavioral effects of partial DNMT3A inactivation have not been systematically assessed. We therefore carried out growth, behavioral, and molecular analyses of mice carrying a constitutive heterozygous deletion of exon 19 of *Dnmt3a* (see *methods*) (Kaneda et al., 2004). We find that this mutation leads to 50% reduction of RNA and protein expression, allowing us to study the *in vivo* effects of heterozygous null mutation of DNMT3A (referred to as $DNMT3A^{KO/+}$) (Figure 3.9A-C).

We first examined phenotypes relevant to overgrowth in individuals with heterozygous DNMT3A mutations (Tatton-Brown et al., 2018), including enlarged body size and obesity (body weight), tall stature (long-bone length), and macrocephaly (skull dimensions).

DNMT3A$^{KO/+}$ mice showed similar body weight to controls in the early postnatal period but were significantly heavier than controls as mature adults (Figure 3.3A). This phenotype mimics a maturity-associated trend toward increasing body weight observed in TBRS patients (Tatton-Brown et al., 2018). Measurements of leg bone length, a major determinant of height in humans (Duyar and Pelin, 2003), indicated a significant increase in total length of long leg bones, with significantly increased tibia length in DNMT3A$^{KO/+}$ mice and a trend towards longer femur length (Figure 3.3B, Figure 3.10A-D). Morphometric analyses of the cranium and mandible indicated very subtle differences between DNMT3A$^{KO/+}$ mice and their wild-type littermates (Figure 3.10E-G). One linear distance spanning the rostrocaudal length of the interparietal bone is larger in DNMT3A$^{KO/+}$ mice relative to wild-type littermates. Two linear distances in the facial region were significantly larger in wild-type mice, while all other comparisons were not significantly different (Figure 3.10E). This result suggests very slight disruptions in growth of the facial region in DNMT3A $^{KO/+}$ mice. Together these findings uncover effects on long bone length that mirror aspects of the human disorder, while skull development in DNMT3A$^{KO/+}$ mice shows more limited effects. Additionally, enlarged body mass in these mice appears to mimic overgrowth and obesity detected in individuals with TBRS (Tatton-Brown et al., 2014, 2018).

To examine neurological and behavioral phenotypes in DNMT3A$^{KO/+}$ mice, we first assessed basic measures of sensation and motor performance such as balance (ledge test, platform test), grip strength (inverted screen test), motor coordination (walking initiation, rotarod), and sensorimotor gating (pre-pulse inhibition). DNMT3A$^{KO/+}$ mice were not significantly different in these assays (Figure 3.11A-G), indicating that heterozygous loss of

DNMT3A does not grossly disrupt sensorimotor function. This allowed us to accurately assess more complex aspects of behavior and cognition.

We next carried out a panel of assays with relevance to neuropathology observed in humans with DNMT3A mutations, including anxiety, autism, and intellectual disability. DNMT3A$^{KO/+}$ mice displayed reduced exploratory behavior during open field testing, including reduced distance traveled and rearing (Figure 3.3C,D). DNMT3A$^{KO/+}$ mice also displayed anxiety-like behavior in this assay, as they spent less time in the center of the open field arena (Figure 3.3E). In tests of climbing behavior, DNMT3A$^{KO/+}$ mice showed longer latency to climb to the bottom of a pole and to the top of mesh screens (Figure 3.11H-J), suggesting that DNMT3A$^{KO/+}$ mice display differences in volitional movement. To further assess anxiety-like behavior, we tested mice in the elevated plus maze and observed that DNMT3A$^{KO/+}$ mice spent less time exploring the open arms of the maze with no change in percent entries made into all arms (Figure 3.3F, Figure 3.11K). Overall, these results demonstrate that the DNMT3A$^{KO/+}$ mice display changes in exploratory behavior, suggesting altered emotionality and increased anxiety-like behaviors.

Intellectual disability is a highly penetrant phenotype observed in patients with DNMT3A mutations, so we assessed learning and memory in the DNMT3A$^{KO/+}$ mice using fear conditioning and Morris water maze tests. In the foot-shock induced fear conditioning task, DNMT3A$^{KO/+}$ mutants showed heightened freezing response during training, as well as contextual and auditory recall phases of conditioned fear testing (Figure 3.3G-I, Figure 3.11L,M). Mutant mice also showed delayed extinction of freezing behavior in response to the auditory cue alone (Figure 3.11M). These alterations in responses do not arise from altered

sensation of pain from the shock (Figure 3.11N), and therefore suggest altered emotionality or cognition in the mutant mice. Assessment of spatial and contextual memory by Morris water maze testing demonstrated that DNMT3A$^{KO/+}$ mice were slower to learn to find a visible platform and did not learn the location of the hidden platform over time to the level of wild-type controls (Figure 3.11O,P), with no differences in swimming speed (Figure 3.11Q,R). There were no significant effects on distance traveled in target zone or platform crossings in the probe trial, though DNMT3A$^{KO/+}$ mice trended towards fewer platform crossings (Figure 3.11S,T). These findings suggest that DNMT3A$^{KO/+}$ mutants do not show severe deficits in learning and memory but do display differences in task performance that further suggest altered emotionality or cognition in these mice.

DNMT3A has been identified as a high confidence autism gene (Sanders et al., 2015) and there is a high rate of autism diagnoses in patients with TBRS (Tatton-Brown et al., 2018). We therefore sought to characterize common phenotypes examined in mouse models of autism (marble burying, three chamber social approach, ultrasonic vocalizations). We detected a significant reduction in marble burying activity for DNMT3A$^{KO/+}$ mice, indicating alterations in repetitive digging behavior (Figure 3.3J, Figure 3.11U). Evaluation of social interaction behaviors in the three-chamber social approach assay for adult mice (Yang et al., 2011) revealed that DNMT3A$^{KO/+}$ mice spend reduced time investigating both mice and objects. They also display reduced overall activity, with no change in preference index for sociability or novelty (Figure 3.12A-D) (Nygaard et al., 2019). Given that carrying out these tests under normal lighting and experimental conditions may induce anxiety responses in addition to probing sociability, we considered that the reduced exploration seen in DNMT3A$^{KO/+}$ may reflect the

anxiety-like phenotypes of these animals rather than changes in sociability. Indeed, when we carried out three-chamber analysis under low lighting conditions to mitigate anxiety in the mice, we detected similar exploratory behavior, as well as similar sociability and social novelty seeking responses in wild-type and DNMT3A$^{KO/+}$ mice (Figure 3.12E-H) (Manno et al., 2020). To explore changes in pro-communication behaviors, we measured maternal-isolation induced ultrasonic vocalizations (Barnes et al., 2017). We detected a reduction in the number of calls in DNMT3A$^{KO/+}$ mice at postnatal day five, suggesting deficits in early pro-social behaviors or developmental delay in the normal acquisition of this behavior (Figure 3.3K). Together these results indicate alterations in some behaviors commonly assessed in mouse models of autism (Chang et al., 2017; Kazdoba et al., 2014; Mei et al., 2016; Simola and Granon, 2019; Takumi et al., 2019), with our findings suggesting a reduction in activity and exploration, as well as changes in communication behaviors.

Overall, our physiological and behavioral analyses demonstrate that heterozygous deletion of DNMT3A results in growth and behavior changes in mice with relevance to the overgrowth, anxiety, communication, and memory-associated phenotypes observed in patients with DNMT3A mutations. These data support a model in which DNMT3A haploinsufficiency contributes to altered growth and behavioral circuits to drive phenotypes in NDD.

Global disruption of DNA methylation in the DNMT3A$^{KO/+}$ brain

We next investigated the epigenomic defects that may underlie the altered behaviors observed in DNMT3A$^{KO/+}$ mice. We first used sparse whole-genome bisulfite sequencing to efficiently survey effects on global DNA methylation levels for multiple brain regions and liver

tissue isolated from wild type and DNMT3A$^{KO/+}$ mice. This analysis detected limited reductions in genome-wide mCG levels in the DNMT3A$^{KO/+}$ brain that were not apparent in non-neural tissue (liver, Figure 3.4A). In contrast, mCA levels were reduced by 30-50% across all brain regions examined in DNMT3A$^{KO/+}$ mice (Figure 3.4B). Analysis of DNA methylation across postnatal development in the cerebral cortex, a brain region with high levels of mCA at 8 weeks, suggested that deficits in mCA appear during initial accumulation of this methyl mark at 1-6 weeks (Figure 3.4B). Thus, global mCA levels in the brain appear to be highly sensitive to heterozygous DNMT3A disruption, while overall global mCG levels are largely maintained.

DNA methylation at specific genomic elements, including promoters, enhancers, and gene bodies is thought to play an important role in regulating transcription. Alterations in methylation at these regions can impact gene expression to affect the development and function of the brain (Clemens et al., 2019; Nord and West, 2019; Stroud et al., 2017). We therefore assessed changes in methylation at base-pair resolution by high-depth whole-genome bisulfite sequencing to identify potential changes in mCA and mCG at these important regulatory sites. For this analysis we focused on the cerebral cortex, as this region is enriched for mCA (Figure 3.4B) and disrupted in ASD and MeCP2 disorders (Clemens et al., 2019; Satterstrom et al., 2019; Sceniak et al., 2016; de la Torre-Ubieta et al., 2016). High-resolution analysis of mCG confirmed the extremely subtle reduction in mCG across all classes of genomic elements (Figure 3.5A-D). We considered however, that CG dinucleotides in specific sites in the neuronal genome may be more sensitive to a partial reduction in DNMT3A activity. For example, in the hematopoietic system, heterozygous disruption of DNMT3A leads to reductions in DNA methylation in genomic regions that can be identified as sensitive to complete loss of DNMT3A (Cole et al.,

2017). We therefore evaluated CG-differentially methylated regions (CG-DMRs) previously identified as becoming highly methylated in the adult cortex (Figure 3.13A) (Lister et al., 2013). Because DNMT3A is the only *de novo* methyltransferase expressed in the postnatal brain, we hypothesized that adult-specific CG-DMRs might be sensitive to a reduction in enzyme activity. Indeed, we found that these sites build up mCG during postnatal development and do not become methylated in a brain-specific DNMT3A conditional knockout mouse (DNMT3A *Nestin*-cKO) (Stroud et al., 2017) (Figure 3.13B). Analysis of these adult-specific CG-DMRs in the DNMT3A$^{KO/+}$ model indicated that these sites are particularly sensitive to partial inactivation of DNMT3A (Figure 3.5A,C,D).

Because a substantial percentage of mCG in neurons can occur in an oxidized, hydroxymethyl form (hmCG), we performed oxidative bisulfite sequencing analysis of DNA from the cortex. This analysis revealed no clear evidence of widespread differential effects on the oxidized or unoxidized forms of mCG across classes of genomic regions in the DNMT3A$^{KO/+}$ (Figure 3.13C), suggesting that hmCG is largely preserved in the DNMT3A$^{KO/+}$ brain. However, future experiments will be required to uncover if there are more subtle site-specific changes in hmCG.

To further search for local sites of altered mCG in the DNMT3A$^{KO/+}$, we performed *de novo* calling of CG-differentially methylated regions using the BSmooth algorithm (Hansen et al., 2012). We identified 6164 hypo- and 378 hyper-CG-DMRs across the genome that met high stringency filters for size and reproducibility (Figure 3.5A-C,E, Figure 3.13D, see *methods*). The hypo-DMRs significantly overlap with the previously identified adult-specific CG-DMRs (Lister et al., 2013) (Figure 3.5F, Figure 3.13E), further supporting the idea that DNMT3A is

haploinsufficient for postnatal mCG deposition at these sites. Examination of the genomic distribution of all DNMT3A$^{KO/+}$ CG-DMRs revealed significant overlap of hypo-DMRs with putative enhancer regions, gene bodies, and promoters (Figure 3.5A-C,F, Figure 3.13E). DMRs were also highly enriched for overlap with CpG island shores, regions disrupted in studies of DNMT3A mutation outside of the nervous system (Cole et al., 2017; Spencer et al., 2017) (Figure 3.5C, Figure 3.13E). Together these findings indicate that a small subset of mCG sites are particularly sensitive to heterozygous loss of DNMT3A and that localization of these CG-DMRs to regulatory elements suggests that these methylation changes could impact gene expression.

We next examined the profile of mCA at genomic elements of different scales that have relevance to gene regulation. In contrast to the limited mCG changes in the DNMT3A$^{KO/+}$, analysis of mCA levels detected consistent 30-50% reductions at nearly all genomic regions examined (Figure 3.5G). This was true of gene bodies, promoters, and CpG island shores. CpG islands show very low mCA levels in wild-type cortex and displayed less reduction of mCA, possibly due to floor effects in bisulfite-sequencing (see *methods*). Comparing the mCA changes within each class of genomic elements as a function of wild-type mCA levels suggested that consistent reductions occurred across the genome, independent of the normal levels of mCA (Figure 3.13F). This result suggests that changes in mCA levels in the DNMT3A$^{KO/+}$ do not preferentially impact specific classes of genomic elements, nor are they more severe in some regions based on the level of mCA normally deposited.

Recent analysis has demonstrated that topologically-associating domains (TADs) of chromatin folding can impact gene regulation by influencing DNMT3A binding and subsequent

mCA deposition (Boxer et al., 2019; Clemens et al., 2019; Nothjunge et al., 2017). The average level of mCA within TADs is associated with the level of mCA at enhancers within TADs, and high-mCA enhancers found in high-mCA TADs are particularly robust targets of repression by MeCP2 (Clemens et al., 2019). We therefore specifically assessed mCA levels at TADs and enhancers genome-wide. This analysis detected reductions in TAD mCA levels that were similar to global reductions in mCA at other genomic elements (Figure 3.5G). Enhancers also showed this pervasive depletion of mCA (Figure 3.5G). Thus, widespread loss of mCA for TADs, and the enhancer elements contained within them, occurs in DNMT3A$^{KO/+}$ mice and has the potential to impact epigenetic control of regulatory elements by MeCP2.

Enhancer dysregulation results from methylation deficits in DNMT3A$^{KO/+}$ mice

We next examined how disruption of DNA methylation can affect epigenetic regulation in DNMT3A$^{KO/+}$ neurons to alter gene expression and disrupt nervous system function. Recent analyses indicate that mCA serves as a binding site for MeCP2 to mediate neuron-specific gene regulation, in part by controlling the activity of distal regulatory enhancer elements (Boxer et al., 2019; Clemens et al., 2019). Loss of MeCP2 in mice leads to genome-wide upregulation of the activating mark Histone H3 lysine 27 acetylation (H3K27ac) at enhancers that contain high levels of mCA and mCG sites, while overexpression of MeCP2 leads to reciprocal downregulation of highly methylated sites. Alterations in enhancer activity in MeCP2 mutants are linked to dysregulation of genes that can then drive nervous system dysfunction (Clemens et al., 2019). These findings suggest that reduced CA methylation in the DNMT3A$^{KO/+}$ would remove binding sites for MeCP2 within enhancers, which could then result in dysregulation of

enhancer activity that partially phenocopies the effects we have observed in MeCP2 mutant mice.

To investigate this possibility directly, we quantified the change in DNMT3A$^{KO/+}$ mCA binding sites for enhancers significantly repressed by MeCP2 (Clemens et al., 2019). These enhancers contain a large number of mCA sites due to high mCA/CA levels and an enrichment of CA dinucleotides within these sequences. As a result, we found that the global 30-50% reduction of mCA in the DNMT3A$^{KO/+}$ leads to a larger loss in the total number of mCA sites at MeCP2-repressed enhancers than at other enhancers genome-wide (Figure 6A,B). Thus MeCP2-repressed enhancers are particularly susceptible to mCA binding site loss from heterozygous mutation of DNMT3A.

To determine if the reduction of mCA sites at MeCP2-repressed enhancers affects their activity, we assessed changes in enhancer activation level by H3K27ac ChIP-seq analysis of the DNMT3A$^{KO/+}$ and wild-type cerebral cortex. This analysis revealed significant changes in acetylation at MeCP2-repressed enhancers (Figure 3.6A,C). Consistent with these effects arising from 30-50% loss of the mCA that normally builds up post-mitotically at enhancers, we detect changes that are concordant with, but smaller than, effects on enhancers previously observed upon complete loss of post-mitotic mCA in the neuron-specific DNMT3A conditional knockout mouse (DNMT3A *Baf53b*-cKO) (Clemens et al., 2019) (Figure 3.6C). We also find that enhancers within highly methylated TADs show significantly higher increases in acetylation compared with enhancers in other TADs (Figure 3.13G). This finding further supports a model in which high mCA levels in TADs leads to repression of enhancers within these TADs by mCA and MeCP2.

Although significantly dysregulated enhancers can be detected in MeCP2 mutants, broad sub-significance-threshold effects also occur genome-wide upon MeCP2 mutation, with enhancers across the genome undergoing dysregulation that is proportional to the number of mC binding sites at these regions (Figure 3.6D) (Clemens et al., 2019). Analysis of H3K27ac changes at enhancers based on the normal density of mCA sites in these sequences genome-wide revealed broad mCA-associated derepression of enhancers in DNMT3A$^{KO/+}$ mouse cortex that is similar to, but smaller in magnitude than, the effects observed in DNMT3A *Baf53b*-cKO and MeCP2 knockout mice (MeCP2 KO). These effects are also reciprocal to effects observed in MeCP2 overexpression mice (MeCP2 OE). Consistent with the limited disruption of mCG genome wide in the DNMT3A$^{KO/+}$ mice, there was only a limited association between changes in enhancer activity and the level of mCG. This limited association contrasts with MeCP2 mutants, where loss of protein binding at both mCG and mCA sites leads to enhancer dysregulation that is associated with both mCA and mCG (Clemens et al., 2019) (Figure 3.6D).

Though there is limited evidence that a global change in mCG impacts enhancer activity in the DNMT3A$^{KO/+}$ cortex, we considered the potential for site-specific changes in mCG (DMRs) to have functional effects on enhancers. We therefore assessed changes in acetylation at enhancers containing CG-DMRs in the DNMT3A$^{KO/+}$ cortex. Indeed, enhancers containing hypo-CG-DMRs display subtle, but significant increases in acetylation compared to enhancers without a DMR (Figure 3.13H). The results thus provide evidence that site specific mCG changes in the DNMT3A$^{KO/+}$ may result in functional consequences.

Together, our epigenomic analyses demonstrate that loss of half of the normal mCA sites in the DNMT3A$^{KO/+}$ cortex results in enhancer dysregulation that overlaps with MeCP2 mutant

mice. On a more limited scale, loss of CG methylation at specific enhancers can also impact enhancer activity. These findings therefore uncover a role for shared neuronal chromatin pathology between DNMT3A and MeCP2 disorders.

Overlapping transcriptional pathology between DNMT3A$^{KO/+}$, MeCP2 disorders, and ASD

The epigenetic alterations we observe in DNMT3A$^{KO/+}$ cerebral cortex can have direct consequences on gene expression to drive neurological dysfunction in mice. Furthermore, the overlapping effects on enhancers that we observe between DNMT3A$^{KO/+}$ and MeCP2 mutant mice suggest that there may be shared transcriptional pathology occurring upon loss of mCA in DNMT3A disorders, as well as through disruption of MeCP2 in Rett syndrome and MeCP2-duplication syndrome. We therefore assessed changes in gene expression in DNMT3A$^{KO/+}$ mice, interrogating the extent to which these effects overlap with those observed upon complete disruption of mCA in the DNMT3A *Baf53b*-cKO and those observed in MeCP2 mutant mice. RNA-seq of DNMT3A$^{KO/+}$ cerebral cortex identified subtle changes in gene expression that are consistent in magnitude with effects observed in other heterozygous NDD models (Fazel Darbandi et al., 2018; Gompers et al., 2017; Katayama et al., 2016) (Figure 3.14A). Additional analysis of noncoding RNA showed no change across major classes of noncoding RNAs in the DNMT3A$^{KO/+}$ cortex (Figure 3.14B).

Gene Ontology analysis of gene expression changes detected terms associated with neurodevelopment processes, such as axonal guidance and recognition, that occur just before or during the period when DNMT3A deposits mCA in the brain (Figure 3.14C). This is consistent with the proposed role of DNMT3A in regulating transcriptional programs as neurons mature.

Additionally, we see significantly reduced expression of relevant ASD genes, Shank2 and Shank3, that suggest dysfunction at the synaptic level (Supplementary Table 3) (Guilmatre, Huguet, Delorme, and Bourgeron, 2014). Our transcriptomic results suggest that genes critical to the formation of neuronal circuits during development are affected upon heterozygous disruption of DNMT3A, with disrupted development resulting in behavioral alterations we observe.

To determine the extent to which transcriptional dysregulation upon disruption of DNMT3A and MeCP2 result in shared effects, we examined the degree to which genes detected as significantly dysregulated in each mutant line overlap with the changes in the DNMT3A$^{KO/+}$ cortex. Indeed, the significantly dysregulated genes in the DNMT3A$^{KO/+}$ overlapped extensively with genes identified as significantly dysregulated in DNMT3A *Baf53b*-cKO and MeCP2 mutant mice (Figure 3.6E) (Clemens et al., 2019). Notably, shared gene expression effects between these mouse models may have important functional impacts. For example, all three deletion models show upregulation of Shroom3, which is associated with developmental defects when mutated in humans (Das et al., 2014; Deshwar et al., 2020), and latrophilin-2, a post-synaptic adhesion molecule with roles in synapse targeting and assembly in multiple brain regions (Anderson et al., 2017; Zhang et al., 2020). The shared dysregulation of genes like these in DNMT3A and MeCP2 disorders, may contribute to synaptic dysfunction and disease pathology.

While a limited gene set is detected as significantly dysregulated in the DNMT3A$^{KO/+}$, we considered if genome-wide alterations in enhancer activity could lead to widespread, subtle dysregulation of gene expression that is below the significance threshold of detection for individual genes. In this way, the transcriptional pathology in the DNMT3A$^{KO/+}$ brain would overlap with similar subthreshold genome-wide effects observed upon loss of neuronal mCA

(DNMT3A *Baf53b*-cKO) and in models of Rett syndrome (MeCP2 KO) and ASD (MeCP2 OE) (Clemens et al., 2019; Gabel et al., 2015). Importantly, evidence from MeCP2 mutants suggests that disease pathology arises from thousands of small changes in gene expression rather than drastic changes in a few significant genes (Chen et al., 2015; Ip et al., 2018; Kinde et al., 2016). As such, we may expect that disruption of the neuronal methylome in DNMT3A$^{KO/+}$ may also lead to numerous changes in gene expression that escape statistical significance yet still contribute to disease pathology. To more comprehensively assess the degree to which subtle, transcriptome-wide changes in the DNMT3A$^{KO/+}$ phenocopy these MeCP2 mutant and DNMT3A *Baf53b*-cKO models, we performed Generally Applicable Gene-set Enrichment (GAGE) analysis (Luo et al., 2009) of genes detected as dysregulated. GAGE analysis revealed highly significant, concordant changes in gene expression in the DNMT3A$^{KO/+}$ for genes identified as dysregulated upon loss of mCA in the DNMT3A *Baf53b*-cKO and in MeCP2 mutants (Figure 3.6F).

To further explore a model in which loss of mCA and enhancer dysregulation drives gene expression changes in the DNMT3A$^{KO/+}$, we examined enhancers associated with significantly upregulated genes. This analysis revealed that these enhancers are enriched for mCA, show a pronounced loss of mCA sites in the DNMT3A mutant, and display significant increases in enhancer acetylation compared to all other enhancers (Figure 3.14D,E). Analysis of genes containing enhancers that harbor a hypo-CG-DMR also detected significant upregulation in gene expression compared to all other genes (Figure 3.14F), suggesting that local changes in mCG may also contribute to transcriptomic changes in DNMT3A$^{KO/+}$ mice.

Having detected overlap in transcriptomic pathology between models of DNMT3A and MeCP2 disorders, we explored if shared gene expression signatures in the DNMT3A$^{KO/+}$ mice extend to models of disorders that do not have established mechanistic links to DNMT3A disorders. We therefore tested if DNMT3A$^{KO/+}$ mice show significant alterations in gene sets identified as dysregulated in other mouse models of NDD and in human gene sets implicated as altered in the autistic brain. GAGE analysis across multiple datasets detected highly significant dysregulation of gene sets identified in CHD8 and PTEN mouse models of overgrowth and ASD (Gompers et al., 2017; Katayama et al., 2016; Tilot et al., 2016) as well as the SETD5 model of NDD (Sessa et al., 2019) (Figure 3.7A). These findings support a role for overlapping gene dysregulation underlying common symptomology found in affected individuals carrying mutations in distinct epigenomic regulatory genes. We further extended our analysis to human gene sets dysregulated in ASD postmortem brains (Gandal et al., 2018; Voineagu et al., 2011). This analysis detected significant changes in the DNMT3A$^{KO/+}$ cortex (Figure 3.7B), and indicated upregulation of candidate genes linked to ASD from human genetics studies (Abrahams et al., 2013; Banerjee-Basu and Packer, 2010) (Figure 3.7B). In addition, analysis of co-expression modules of human brain development (Parikshak et al., 2013) showed overlap with several neurodevelopmental modules including those that increase during early cortical development and are enriched for ASD risk genes (M13, M16, and M17) (Figure 3.14G). Modules involved in regulation of nucleic acids and gene regulation that are expressed early in development and decrease over time are also increased upon heterozygous loss of DNMT3A (M2 and M3) (Figure 3.14G). These results indicate that important sets of genes with opposing developmental trajectories and function are altered upon loss of DNMT3A regulation.

Resampling analysis indicates that significant dysregulation of these mouse and human gene sets is not driven by enriched expression of these genes in the cortex (Figure 3.14H). Together these findings suggest that the DNMT3A$^{KO/+}$ mouse shares overlapping transcriptional pathology with gene expression changes underlying ASD.

3.4 Discussion

Our functional analysis of NDD-associated DNMT3A mutations together with our *in vivo* studies provide an initial working model of molecular etiology in DNMT3A disorders. Diverse *de novo* missense mutations that arise in affected individuals disrupt enzyme function by disabling the capacity of the enzyme to localize to chromatin in the nucleus, altering the ADD-regulatory domain, or disrupting the activity of the methyltransferase domain. Loss-of-function effects resulting from these missense mutations, or early truncations and gene deletions in other individuals, lead to insufficient DNMT3A activity. The limited DNMT3A activity in turn causes deficits in the deposition of mCG at specific sites during development, as well as a massive deficit in postnatal mCA accumulation throughout the brain. These changes in DNA methylation lead to alterations in epigenomic regulation, including subtle but widespread disruption of mCA-MeCP2-mediated enhancer regulation in adult neurons, resulting in gene expression changes that can contribute to deficits in nervous system function.

Our studies of DNMT3A mutations not only provide insight into the molecular etiology of DNMT3A disorders, but also serve as a model for understanding the functional effects of diverse *de novo* mutations underlying neurodevelopmental disorders. Exome sequencing studies have identified a large and growing list of mutations in genes encoding epigenetic regulators in

individuals with NDD. Many of these are missense mutations and occur as heterozygous disruptions (McRae et al., 2017; Satterstrom et al., 2019), leaving it unclear if simple loss-of-function effects are sufficient to drive pathology through haploinsufficiency, or if more complex effects play a role when individual amino acids are altered. In addition, while identification of multiple mutations in a gene can implicate disruption of the gene as causative for NDD, it remains possible that a subset of the mutations identified in affected individuals, particularly missense mutations, are not in fact deleterious or causative. Functional testing of these variants is therefore necessary to determine whether they underlie disease. Here, our analysis of DNMT3A mutations in multiple functional assays has uncovered diverse mechanisms by which the protein can be disrupted while pointing to a shared loss of function in the deposition of neuronal DNA methylation. Notably, it is only by assessing multiple aspects of protein function (i.e. expression, localization, activity, and cellular mCA levels) that we can detect deficits for each mutation tested. For example, mutation of the ADD domain only disrupts deposition of mCA in cells, possibly due to loss of regulation that can only be assessed in the endogenous chromatin context. Together our findings establish deleterious effects of diverse DNMT3A mutations and underscore the importance of multidimensional analysis of *de novo* mutations to fully assess their role in NDD.

Our *in vivo* analyses show that heterozygous deletion of DNMT3A mirrors multiple key features of DNMT3A disorders, including tall stature (increased long bone length), increased body weight, and behavioral alterations. Detection of robust anxiety-like phenotypes in multiple assays, deficits in pro-social communication, and alterations in repetitive behaviors align with observed human phenotypes. In contrast, lack of severe deficits in learning and memory assays

in our mouse model may indicate that some regions and systems in humans are more susceptible to DNMT3A disruption than in mice. However, we do detect subtle alterations in behavior in these assays and the lack of strong deficits may also reflect insensitivity of the methods used in mice to measure specific aspects of disrupted cognition.

In all, our *in vivo* analysis indicates that heterozygous deletion of DNMT3A results in effects which can guide future studies of molecular, cellular, and organismal dysfunction caused by mutation of DNMT3A. We employed the DNMT3A$^{KO/+}$ mouse experimental system to assess how heterozygous DNMT3A disruption impacts epigenetic regulation in the brain. Our analysis of DNA methylation in tissues from DNMT3A$^{KO/+}$ mice detected subtle changes in genome-wide mCG levels across brain regions, with no global mCG effects in non-neural tissue. Analysis of local changes in mCG in the brain detected evidence of disrupted CG methylation at sites methylated during postnatal development (i.e. adult hyper CG-DMRs) and at regulatory elements including enhancers. These effects have the potential to disrupt gene expression and contribute to neurological alterations in these mice. The limited nature of mCG effects is likely due to the redundant function of the other DNA methyltransferases. The maintenance methyltransferase DNMT1 has the capacity to preserve existing mCG patterns during cell divisions (Jeltsch et al., 2018). In addition, the *de novo* methyltransferase DNMT3B is expressed with DNMT3A in many tissues during early development and could provide critical redundancy for mCG patterning (Okano et al., 1999). Nonetheless, some site-specific changes in mCG are also likely to occur in early development and in non-neural tissues. For example, constitutive heterozygous deletion of DNMT3A has been shown to disrupt mCG patterns in the blood and alter

hematopoietic lineages (Cole et al., 2017). These changes in mCG can contribute to changes in growth and other phenotypes observed in mice and humans.

In contrast to mCG, we detect a global reduction in mCA to approximately 30-50% of wild-type levels in DNMT3A$^{KO/+}$ cortex, striatum, cerebellum, and hippocampus. These results generalize and extend findings in the hypothalamus (Sendžikaitė et al., 2019), demonstrating the susceptibility of broad neuronal types and circuits to heterozygous loss of DNMT3A. The susceptibility of mCA to heterozygous loss of DNMT3A is likely due to several related factors. For example, DNMT3B is not expressed in postnatal neurons (Lister et al., 2013), and DNMT1 is not capable of depositing mCA (Jeltsch et al., 2018), making all mCA build-up in neurons dependent on DNMT3A. In addition, the enzyme has slow kinetics for activity on CA sites (Zhang et al., 2018) and deposition of mCA genome-wide by DNMT3A must take place in a restricted time window (1-6 weeks) when the enzyme is highly expressed and active in neurons (Clemens et al., 2019; Lister et al., 2013; Stroud et al., 2017). These constraints may make enzyme levels limiting for mCA accumulation in neurons, providing an explanation for why global mCA in the brain is sensitive to DNMT3A gene dosage. Notably, our findings suggest that manipulations that activate the remaining DNMT3A, or prolong its high early postnatal expression, might rescue deficits in mCA deposition. Conversely, duplication of the DNMT3A gene could result in too much deposition of mCA and possibly cause significant neural dysfunction akin to those effects seen in MeCP2 duplication disorder. Future studies can assess the feasibility of rescue approaches and explore if DNMT3A duplication alters brain function.

Our analysis of chromatin changes downstream of altered DNA methylation has uncovered a striking point of shared molecular disruption across models of DNMT3A disorders,

Rett syndrome, and MeCP2 duplication syndrome. While the clinical profile and pathophysiology of DNMT3A disorders is clearly distinct from MeCP2 disorders, we have shown here that loss of approximately a quarter of MeCP2 binding sites across the neuronal genome in the DNMT3A$^{KO/+}$ cortex results in subtle but widespread disruption of mCA-associated enhancer regulation that partially phenocopies loss of MeCP2. This enhancer dysregulation can be linked to shared alterations in gene expression across these models. Given the critical roles of MeCP2-regulated genes for nervous system function (Gabel et al., 2015; Lagger et al., 2017; Lyst and Bird, 2015), these overlapping epigenomic and transcriptomic effects likely contribute to aspects of neurologic dysfunction observed in DNMT3A disorders.

Overall, we find that the behavioral effects observed in the DNMT3A$^{KO/+}$ mice are less severe than homozygous conditional DNMT3A knockout and MeCP2 knockout mice, which present with decreases in motor coordination, severely reduced body size, and increased mortality (Nguyen et al., 2007; Stroud et al., 2017; Tillotson et al., 2017). These behavioral differences between models mirror the differences in effects on mCA deposition and readout in these strains, where DNMT3A conditional knockout mice lose all neuronal mCA and MeCP2 KO mice lose a major reader of both mCA and mCG, while the DNMT3$^{KO/+}$ mice lose only half of the mCA sites and a limited set of localized mCG sites. The persistence of many mCA and mCG binding sites for MeCP2 in the DNMT3A$^{KO/+}$ may partially explain how DNMT3A mutations in humans manifest with less severe symptomology than in Rett Syndrome. Independently, absence of DNMT3A early in prenatal development can contribute to overgrowth and other non-overlapping aspects of DNMT3A and MeCP2 disorders. Together, our findings show that disruption of mCA-MeCP2 mediated enhancer regulation likely contributes to multiple

disorders with distinct symptomology, defining a site of convergent molecular etiology underlying heterogeneous clinical syndromes.

Our transcriptomic analysis of changes of ASD/NDD gene sets in DNMT3A mice has further detected overlap with NDD beyond MeCP2 disorders, including both mouse models of NDD/ASD (e.g. CHD8) and gene sets identified in human idiopathic ASD. As additional transcriptomic studies of mouse models and human NDD brain emerge, systematic analyses of gene expression effects can identify shared aspects of transcriptional pathology that can contribute to cognitive and social deficits across diverse causes of NDD. Notably, the large number of chromatin-modifying enzymes mutated in these disorders raises the possibility that shared transcriptomic effects emerge from common chromatin pathology. Our study has identified alterations in mCA and enhancer regulation as a potential site of convergent dysfunction in MeCP2 and DNMT3A disorders. Future studies may identify additional gene disruptions in which alterations in mCA and enhancer dysregulation contribute to molecular pathology, expanding the role of "methylopathies" in neurodevelopmental disease.

3.5 Methods

Animal husbandry

All animal protocols were approved by the Institutional Animal Care and Use Committee and the Animal Studies Committee of Washington University in St. Louis, and in accordance with guidelines from the National Institutes of Health (NIH). Mice were housed in a room on a 12:12 hour light/dark cycle, with controlled room temperature (20-22°C) and relative humidity (50%). Home cages measured 28.5 cm x 17.5 cm x 12 cm and were supplied with corncob bedding and standard laboratory chow and water. All mice were group-housed and adequate measures were taken to minimize animal pain or discomfort.

Transgenic animals

Male and female homozygous *Dnmt3a*$^{flx/flx}$ mice (Kaneda et al., 2004) were bred together for viral-mediated DNMT3A replacement assay culture experiments. To generate the DNMT3A heterozygous mouse model, *Dnmt3a*$^{flx/flx}$ mice were crossed to CMV:Cre (B6.C-Tg(CMV-cre)1Cgn/J) to generate *Dnmt3a*$^{KO/+}$Cre:CMV$^{+/-}$ offspring. *Dnmt3a*$^{KO/+}$Cre:CMV$^{+/-}$ progeny were bred to C57BL/6J to outcross the cre recombinase and generate experimental genotype (DNMT3A$^{KO/+}$). Mice were genotyped with ear-DNA by PCR for *Dnmt3a* and *Cre*, and recombination was tested. Subsequent experimental animals were generated from *Dnmt3a*$^{KO/+}$ males mated to C57BL/6J females to generate *Dnmt3a*$^{KO/+}$ and *Dnmt3a*$^{+/+}$ experimental and control animals for experiments. *Dnmt3a*$^{KO/+}$ females were not used for breeding to avoid social differences in mothering from mutant dams. Mice were weighed at a variety of timepoints to assess growth.

Method Details

Immunocytochemistry

Staining. Neuro-2a cells (ATCC, CCL-131) were grown on coverslips and transfected with FLAG-tagged WT or mutant mouse DNMT3A plasmids and GFP plasmid. This FLAG-tag allows for quantification of transfected DNMT3A protein only instead of endogenous DNMT3A. Coverslips were fixed with 4% paraformaldehyde in PBS for 20 minutes, permeabilized with 0.1% Triton X-100 in PBS for 10 minutes, and blocked with 1% BSA in PBS for 1 hour all at room temperature. Coverslips were incubated overnight at 4˚C in anti-DDDDK tag (FLAG-tag) primary antibody (Abcam, 1:5000, ab1162). Coverslips were then washed in PBS and incubated for 1 hour at room temperature with fluorescent secondary antibody (ThermoFisher, 1:500, A-11011) and counterstained with DAPI. *Imaging.* Images were captured using a Nikon A1Rsi confocal microscope with a 20x air objective. Laser settings were kept constant for each image. *Analysis/Quantification.* Cells were counted using an automatic threshold in FIJI and manually classified as displaying nuclear or non-nuclear signal by a blinded observer. This was determined by evaluating the overlap of FLAG signal (DNMT3A) with DAPI signal (nucleus). For mutants that did not reach expression levels comparable to the WT or for images that had too few positive cells, cell number was counted manually. 8 separate transfections were run, with each mutant being counted over 3 or more independent experiments. Sample sizes are as follows: WT, 15 images, 880 cells; W297del, 8 images, 435 cells; I310N, 8 images, 492 cells; S312fs11x, 12 images, 321 cells; G532S, 9 images, 695 cells; M548K, 7 images, 333 cells; V665L, 6 images, 635 cells; Y735C, 16 images, 613 cells; R749C, 8 images, 667 cells; P904L, 7 images, 692 cells. While some mutants showed reduced protein expression, all proteins tested showed adequate signal assessed as nuclear or non-nuclear. Percent nuclear was assessed per image and a generalized linear model was run comparing each mutant to WT. *P* values for each mutant were then Bonferroni corrected. Localization to the nuclear periphery was evident for the WT FLAG-tagged protein and all nuclear-localized mutants, but no consistent differences in this

localization was observed between WT and any mutants. We chose to use a generalized linear model with Bonferroni correction to allow for us to compare ratios of percent nuclear signal while accounting for experimental and biological replicates.

Modeling of DNMT3A disease mutations

HEK293T (ATCC, ACS-4500) or Neuro-2a cells (ATCC, CCL-131) were transfected with GFP and FLAG-tagged WT or mutant mouse DNMT3A plasmids. Collected cell lysates were ruptured by 3 freeze/thaw cycles using liquid nitrogen, or sonication ~42 hours after transfection. Samples were then either used for western blotting (Neuro-2a cells), the *in vitro* radioactive methyltransferase assay (HEK293T cells).

qRT-PCR

RNA isolated from neuronal cortical culture or mouse brain tissue was reverse transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). *Dnmt3a* and *Actb* were measured by qPCR using the Power SYBR™ Green PCR Master Mix and primers *Actb* (F:AAGGCCAACCGTGAAAAGAT, R:GTGGTACGACCAGAGGCATAC) and *Dnmt3a* (F:GGCCTTCTCGACTCCAGATG, R:TTCCTCTTCTCAGCTGGCAC). Relative quantity of *Actb* and *Dnmt3a* cDNA was determined by comparing the Ct of each primer set in each sample to a standard curve and then normalizing the DNMT3A signal by the ACTB signal. We chose to compare experimental conditions to WT samples using Student's T-Tests with Bonferroni correction, as these are two normally distributed groups (visually checked) with similar variability.

***In vitro* radioactive methyltransferase assay**

30 µl of cell lysate from HEK293T cells was used in the methyltransferase reaction previously described (Russler-Germain et al., 2014). Lysates were incubated at 37˚C for 20 hours in 5 µl reaction buffer of 20 mM HEPES, 30 mM NaCl, 0.5 mM DTT, 1 mM EDTA, 0.2 mg/ml BSA, 5 mM 3 H-labeled SAM (PerkinElmer, NET155050UC) and 500 ng/µl Poly(dI-dC) substrate (Sigma P4929). Substrate was purified (Macherey-Nagel NucleoSpin Gel and PCR Clean-up) and radioactivity measured using a scintillation counter. In instances where DNMT3A mutant showed altered protein expression, cell lysate was re-balanced to match protein expression of WT DNMT3A. The activity of endogenous DNMT3A present in the cells was accounted for by normalizing values to GFP alone. Only experimental replicates where WT DNMT3A showed a 1.5-fold increase compared to GFP alone were used for subsequent analysis. Outliers beyond 2 standard deviations above or below the mean were removed. Number of independent replicates are as follows: W297del, 18; I310N, 19; S312fs11x, 4; G532S, 10; M548K, 15; V665L, 11; Y375C, 13; R749C, 7; P904L, 14. Significance was assessed using a one-sample student's t-test, as we are comparing groups normalized to WT and GFP back to the normalized value of 1.

Viral-mediated DNMT3A replacement assay

Functional activity of DNMT3A mutants in cortical neurons was determined by measuring methylation build-up *in vitro*. Timed mating of DNMT3A$^{flx/flx}$ females and DNMT3A$^{flx/flx}$ males was performed to collect embryonic cortical DNMT3A$^{flx/flx}$ neurons at embryonic day 14.5. At E14.5, DNMT3A$^{flx/flx}$ cortical neurons were isolated and plated (DIV 0). On DIV 3, neurons were either not perturbed or virally transduced with one of three conditions: 1) Cre only, 2) Cre and WT DNMT3A, or 3) Cre and mutant DNMT3A. DNA and RNA were isolated on DIV 12.5 using the AllPrep DNA/RNA Kit (Qiagen, 80204). DNA was used for whole genome bisulfite sequencing, and RNA was used for qRT-PCR for DNMT3A. Results of pilot experiments (not shown) indicate that at this timepoint mCA is not saturated and sensitive to total levels of DNMT3A activity. Number of independent replicates are as follows: W297del, 5; S312fs11x, 4; G532S, 7; M548K, 6; V665L, 10; Y375C, 11; R749C, 6; P904L, 8. Buildup of methylation over development was carried out without performing any viral transduction. Significance was assessed using a one-sample student's t-test or 1-way ANOVA.

Ultrasonic vocalization and analysis

A total of 76 DNMT3A$^{KO/+}$ (n=30, 16 male and 14 female) and litter-matched WT (n=46, 25 male and 21 female) mice were used for ultrasonic pup vocalization (USV) recording and analyzed as previously described (Barnes et al., 2017; Dougherty et al., 2013; Holy and Guo, 2005). Dams were removed from the nest for a 10-minute acclimation, and individual pups had their body temperature measured using an infrared laser thermometer. Pups were then removed from their nest and placed in a dark, enclosed chamber. Ultrasonic vocalizations were recorded for 3 minutes with an Avisoft UltraSoundGate CM16 microphone and 416H amplifier using Avisoft Recorder software (gain = 6 dB, 16 bits, sampling rate = 250 kHz). Pups were then weighed and returned to their nest and littermates. All mice were recorded at postnatal days 5, 7, and 9, and on either day 11 or 15. Frequency sonograms were prepared and analyzed in MATLAB (frequency range = 40 kHz to 120 kHz, FFT size = 256, overlap = 50%) with individual syllables identified and counted according to previously published methods (Dougherty et al., 2013; Holy and Guo, 2005). Significance was assessed using a within-subjects repeated measures ANOVA over timepoints 5-9, as these were when there was data from all experimental subjects, and these are optimal testing times where number of calls was highest.

Marble burying

A total of 27 DNMT3A$^{KO/+}$ (n=13, 8 male and 5 female) and litter-matched WT (n=14, 7 male and 7 female) mice were used for marble burying. Marble burying is a natural murine behavior and has been used to indicate repetitive digging as well as anxiety-related behaviors. Protocol was adapted from previously published methods (Lazic, 2015; Maloney et al., 2019a). In brief,

8-week old mice were placed in a transparent enclosure (28.5 cm x 17.5 cm x 12 cm) with clean aspen bedding and 20 dark blue marbles evenly spaced in a 4 x 5 grid on top of the bedding. Animals explored freely for 30 minutes. The number of buried marbles were counted every 5 minutes by two independent blinded observers. Marbles were considered "buried" if they were at least two-thirds covered by bedding. Enclosure and marbles were cleaned thoroughly between animals. Significance was assessed using a within-subjects repeated measures ANOVA to determine if rate of burying marbles is different between genotypes. These statistical methods are more appropriate than a simple t-test at 30 minutes, as mice may have buried all marbles before this timepoint, and significant changes in marble burying behavior may have occurred at earlier timepoints in the assay.

Adult behavioral battery

A total of 72 DNMT3A$^{KO/+}$ (n=39, 18 male and 21 female) and litter-matched WT (n=33, 15 male and 18 female) mice were used for adult behavioral testing. Mice were housed in mixed genotype home cages with 2-5 animals per cage, and all tests were performed during the light cycle. All experimenters were blinded to genotype during testing. For increased experimental rigor and reproducibility, we used three separate cohorts of mice to ensure quality and consistency in any observed phenotypes.

Testing started when mice were 3-4 months of age. The sequence of behavioral testing was designed to minimize carry-over effects across behavioral tests. Most assays were performed on cohorts 1 and 2 with cohort 3 being performed to test for reproducibility in some assays. Because of differences in testing sequences and exposure of mice to prior tests between cohorts, we examined separate cohorts individually and looked at combined cohorts. Testing was performed by the Washington University in St. Louis Animal Behavior Core.

One-hour locomotor activity

Locomotor activity was evaluated by computerized photobeam instrumentation in transparent polystyrene enclosures (47.6 cm x 25.4 cm x 20.6 cm) as previously described (Wozniak et al., 2004). Activity variables such as ambulations and vertical rearings were measured in addition to time spent in a 33 cm x 11 cm central zone.

Sensorimotor battery

Mice were assayed in walking initiation, balance (ledge and platform tests), volitional movement (pole and inclined screens), and strength (inverted screen) as previously described (Grady et al., 2006; Wozniak et al., 2004). For the walking initiation test, mice were placed on the surface in the center of a 21 cm x 21 cm square marked with tape and the time for the mouse to leave the square was recorded. During the balance tests, the time the mouse remained on an elevated plexiglass ledge (0.75 cm wide) or small circular wooden platform (3.0 cm in diameter) was

recorded. During the Pole test, mice were placed at the top of a vertical pole with nose pointing upwards. The time for the mouse to turn and climb down the pole was recorded. For the inclined screen tests, a mouse was placed (oriented head-down) in the middle of an elevated mesh grid measuring 16 squares per 10 cm angled at 60° or 90°. Time for the mouse to turn 180° and climb to the top was recorded. For the inverted screen test, a mouse was placed on a similar screen and when the mouse appeared to have a secure grasp of the screen, the screen was inverted 180° and the latency for the mouse to fall was recorded. All tests had a duration of 60 seconds, except for the pole test which was 120 seconds. Two separate trials were done on subsequent days and averaged time of both trials was used for analysis. Data from the walking initiation, ledge, and platform tests were not normally distributed and therefore analyzed using Mann-Whitney U tests.

Continuous and accelerating rotarod

Motor coordination and balance were assessed using the rotarod test (Rotamex-5, Columbus Instruments, Columbus, OH) with three conditions: a stationary rod (60-second maximum), a rotating rod at constant 5 rpm (60-second maximum), and a rod with accelerating rotational speed (5 – 20 rpm, 180-second maximum) as previously described (Grady et al., 2006). This protocol is designed to minimize learning and instead measure motor coordination, so testing sessions were separated by 4 days to allow for extinction. Testing included one trial on stationary rod, and two trials on both the constant-speed rotarod and accelerating rotarod. Later timepoints in the constant speed rotarod test failed tests of normality, as the majority of mice stayed on the rotating rod for all 60 seconds. However, data were analyzed with two-way repeated-measures ANOVA.

Morris water maze

Spatial learning was assessed as previously described (Wozniak et al., 2004). Cued trials (visible platform, variable location) and place trials (submerged, hidden platform, consistent location) were conducted in which escape path latency, length, and swimming speeds were recorded. Animal tracking was done using a computerized system (ANY-maze, Stoelting). During cued trials, animals underwent 4 trials per day over 2 consecutive days with the platform being moved to a different location for each trial with few distal spatial cues available. Each trial lasted no longer than 60 seconds, with a 30-minute interval between each trial. Performance was analyzed across four blocks of trials (2 trials/block). After a three-day rest period, animals were tested on place trials, in which mice were required to learn the single location of a submerged platform with several salient distal spatial cues. Place trials occurred over 5 consecutive days of training, with 2 blocks of 2 consecutive trials (60-second trial maximum, 30-second inter-trial-interval after the mouse has reached the platform) with each block separated by 2 hours. Mice were released into different quadrants over different trials. Place trials were averaged over each of the five consecutive days (4 trials/block). One hour after the final block, a probe trial occurred

(60-second trial maximum) in which the platform is removed, and the mouse is released from the quadrant opposite where the platform had been located. The time spent in pool quadrants, and the number of crossings over the exact platform location were recorded. DNMT3A$^{KO/+}$ mice showed a small, but significant reduction in target zone time in cohort 2, though there was no difference in cohort 1. Additionally, female mice had significantly faster swimming speeds than male mice across both genotypes.

Elevated plus maze

Anxiety-like behaviors were examined using the elevated plus maze as previously described (Boyle, 2006). The apparatus contains a central platform (5.5 cm x 5.5 cm) with two opposing open arms and two opposing closed arms (each 36 cm x 6.1 cm x 15 cm) constructed of black Plexiglas. Mouse position is measured using beam-breaks from pairs of photocells configured in a 16 x 16 matrix and outputs are recorded using an interface assembly (Kinder Scientific) and analyzed using software (MotoMonitor, Kinder Scientific) to determine time spent, distance traveled, and entries made into open arms, closed arms, and the center area. Test sessions were conducted in a dimly lit room with each session lasting 5 minutes and each mouse tested over 3 consecutive days. Data shown are from day 1. All mice showed a decrease in time, distance, and entries into open arms on days 2 and 3. There was no significant difference between genotypes in percent entries into open arms (Figure 3.11K; P=0.137; unpaired Student's T-Test) or total entries into arms (data not shown), indicating that both genotypes explored the maze. Percent distance traveled in open arms showed similar effects to percent time in open arms (Percent distance traveled: P=0.027; unpaired Student's T-Test). Analysis of these data in individual cohorts detected DNMT3A$^{KO/+}$ significant effects for the percent of open arm time on the first day in cohorts 1 and 3, with no evidence of an effect in cohort 2. Individual cohorts also showed no significant difference between genotypes in percent open arm entries suggesting that mice explored the elevated plus maze sufficiently to detect anxiety-like behaviors.

Acoustic startle/prepulse inhibition

Sensorimotor gating was evaluated as previously described (Dougherty et al., 2013; Gallitano-Mendel et al., 2008; Hartman et al., 2001). In short, mice were presented with an acoustic startle response (120 dB auditory stimulus pulse, 40 ms broadband burst) and a pre-pulse (response to pre-pulse plus startle pulse). Stimulus onset began at 65 seconds, and 1ms force readings were obtained and averaged to produce an animal's startle amplitude. 20 startle trials were presented in 20 minutes. The first 5 minutes were an acclimation period where no stimuli above the 65 dB background were presented. The session started and finished with 5 consecutive startle (120 dB pulse) trials. The middle 10 trials were interspersed with pre-pulse trials, consisting of an additional 30 presentations of 120 dB startle stimuli preceded by pre-pulse stimuli of 4, 12, or 20 dB above background (10 trials for each PPI trial type). To calculate

percent pre-pulse inhibition, we used $\%PPI = 100 \times (ASR_{startle\ pulse\ alone} - ASR_{prepulse + startle\ pulse})/ASR_{startle\ pulse\ alone}$.

Conditioned fear

Fear conditioning was done as previously described (Maloney et al., 2019a, 2019b). Mice were habituated to an acrylic chamber (26 cm x 18 cm x 18 cm) containing a metal grid floor and an odorant and was illuminated by LED light which remained on for the duration of the trial. Day 1 testing lasted 5 minutes in which, after a brief baseline period with no tone, an 80 dB tone sounded for 20 seconds at trial timepoints 100, 160, and 220 seconds. A 1.0 mA shock (unconditioned stimulus) occurred within the last 2 seconds of the tone (conditioned stimulus). Baseline freezing behavior during the first 2 minutes and the freezing behavior during the last 3 minutes was quantified using image analysis (Actimetrics, Evanston, Illinois). On Day 2, testing lasted for 8 minutes in which the light was illuminated but no tones or shocks were presented. On Day 3, testing lasted for 10 minutes in which the mouse was placed in an opaque chamber with a different odorant than the original test chamber. After a brief baseline period with no tone, the 80 dB tone began at 120 seconds and lasted for the remainder of the trial and freezing behavior to the conditioned auditory stimulus was quantified for the remaining 8 minutes. DNMT3A$^{KO/+}$ mice show elevated freezing levels during training and testing, which may reflect a baseline propensity to freeze, stronger fear conditioning, or an emotional hypersensitivity to the foot-shock. Additionally, our data suggest that the increased freezing in the DNMT3A$^{KO/+}$ mice does not appear to be due to an increased pain response, as mutant mice showed similar levels of shock sensitivity. Evaluation of baseline freezing levels in individual cohorts only showed significance in one of the two cohorts tested.

3-Chamber social approach

Sociability was assayed as previously described (Moy et al., 2004; Silverman et al., 2011). Mice were tested in a rectangular all-Plexiglas apparatus (each chamber measuring 19.5 cm x 39 cm x 22cm) divided into three chambers with walls containing rectangular openings (5 cm x 8 cm) and sliding doors. The apparatus was in a room with indirect light and was cleaned between tests with Nolvasan solution. Stimulus mice were contained within a small stainless-steel withholding cage (10 cm height x 10 cm diameter; Galaxy Pencil/Utility Cup, Spectrum Diversified Designs), allowing minimal contact between mice without allowing fighting. Between tests, withholding cages were cleaned with 75% ethanol solution. A digital video camera recorded movement of the mouse within the apparatus and allowed for tracking with ANY-maze (Stoelting). Distance and time spent in each chamber and investigation zones surrounding the withholding cages were recorded. Zones were defined as 12 cm in diameter from the center of withholding cages.

The test sequence consisted of 4 consecutive 10-minute trials in which the test mouse is placed in the middle chamber and allowed to freely explore the environment. In the first trial, the mouse is placed in the middle chamber with the doors to other chambers shut. In the second trial, the mouse is placed in the middle chamber and can explore all three chambers of the task, allowing it to acclimate to the environment. Neither genotype tested showed a preference towards a side of the chamber during this habituation. For the third trial, a sex-matched novel conspecific was placed within a withholding cage with the other cage remaining empty. For the fourth trial, the same sex-matched conspecific was in one withholding cage, while a new unfamiliar sex-matched stimulus mouse was placed in the other withholding cage. The locations of stimuli mice were counterbalanced across groups for the third trial and randomized novel or familiar for the fourth trial.

Reduced anxiety 3-chamber social approach

An add-on cohort was used to investigate sociability in a low-light environment with minimal distractions and stressors. A total of 37 DNMT3A$^{KO/+}$ (n=18, 9 male and 9 female) and litter-matched WT (n=19, 9 male and 10 female) mice 10-12 weeks of age were used. These mice had not been used for previous behavioral analysis. Two additional DNMT3A mice were removed from the cohort prior to analysis, as they appeared sickly and showed reduced movement. Social approach was carried out as described previously (Manno et al., 2020). In brief, mice were tested in a rectangular clear acrylic apparatus (60 cm x 40.5 cm) separated into three equally sized chambers divided by walls with sliding doors (6 cm x 6 cm), and with one cup placed in each of the edge chambers. The apparatus was in an isolated, quiet room with minimal sound and low-light (270 lux). Three 10-minute phases were carried out. In the first phase, the test mouse freely explored all chambers of the apparatus. In the second phase, a sex-matched conspecific mouse was added to the one of the cups in the side chambers, and the test mouse was allowed to freely explore. In the third phase, a sex-matched novel conspecific was added to the remaining empty cup and the test mouse was allowed to explore. Between experimental mice, the apparatus was cleaned with 70% ethanol. A digital video camera recorded the sessions and mouse location was measured. For analysis, only the first 5 minutes of each phase was used, as mice rapidly habituate to this task (Manno et al., 2020).

Statistical analysis for behavioral tests

Behavioral data were analyzed with R v3.3.2 (including the ANOVA function from the Car package in R (Fox and Weisberg, 2011)) and plots were made using GraphPad Prism 7.03a. Normality was assessed using the Shapiro-Wilkes test and visually confirmed. Data not normally distributed were analyzed using non-parametric tests, with the exception of continuous rotarod data. No consistent genotype by sex interaction effects were observed for any tests. As expected, bodyweights were significantly different between males and females and we therefore presented

this data separated by sex. Data were collapsed across sex for presentation across all other tests. Statistical testing was performed using planned assay-specific methods, such as using Student's T-Tests for single parameter comparisons between genotypes, and within-subjects two-way repeated-measures ANOVA for comparisons across timepoints. Individual timepoints within repeated measures tests were evaluated using Sidak's multiple comparisons test. Individual cohorts were analyzed separately and in aggregate with similar trends seen across cohorts (Supplementary Table 5), therefore data from all cohorts were included together.

Tissue collection

Brain tissue was dissected from DNMT3A$^{KO/+}$ and WT littermate mice in ice-cold PBS, flash-frozen in liquid nitrogen, and stored at -80°C.

Western blotting

Western blotting from cell culture. Neuro-2a cells were collected and combined with 2x laemmli buffer with 5% β-mercaptoethanol. Samples were passed through a Wizard Column (Fisher, Wizard Minipreps Mini Columns, PRA7211), boiled for 5 minutes, and run on a BioRad 4-12% acrylamide gel at 125 V for 60 minutes. Samples were then transferred to a nitrocellulose membrane, which was bisected between 37kDa and 50kDa bands. Membranes were blocked with 3% bovine serum albumin in TBS-T for 1 hour at room temperature and then the lower membrane was immunostained with anti-GFP (ThermoFisher, 1:2000, A-11122) while the upper membrane was immunostained with anti-DDDDK (Abcam, 1:1000, ab1162) for 12-16 hours at 4°C. All primary and secondary antibodies were diluted in 3% Bovine Serum Albumin in TBS-T. Membranes were next washed with TBS-T, and then incubated for 1 hour at room temperature with IR-dye secondary antibody (IRDye 800CW Donkey anti-Rabbit, LI-COR Biosciences, 1:15,000, product number: 926-32213). Blots were then washed with PBS, imaged using the LiCOR Odyssey XCL system, and quantified using Image Studio Lite software (LI-COR Biosciences). FLAG (DDDDK) and GFP levels were normalized to a standard curve, and protein levels are expressed as normalized DDDDK values divided by normalized GFP values to enable comparison of FLAG (DDDDK) levels between blots. Each blot included a standard curve and WT samples. Outliers beyond 2 standard deviations above or below the mean were removed. Number of independent replicates are as follows: WT, 29; W297del, 7; I310N, 7; S312fs11x, 12; G532S, 7; M548K, 9; V665L, 7; Y375C, 8; R749C, 6; P904L, 7. Significance was assessed using a one sample Student's T-Test, as protein expression levels were normalized to GFP and WT, and mutant protein expression was compared to the normalized WT value of 1.

Western blotting from tissue. Brain tissue samples were homogenized with a dounce homogenizer in buffer with protease inhibitors (10mM HEPES pH 7.9, 10mM KCl, 1.5mM $MgCl_2$, 1mM DTT, 10mM EDTA). A portion of the lysate was removed and 1% SDS was

added. Samples were boiled for 10 minutes, followed by a 10-minute spin at 15,000g. Supernatant was collected and run through a Wizard Column (Fisher, Wizard Minipreps Mini Columns, PRA7211), then diluted in LDS sample buffer with 5% β-mercaptoethanol. Samples were boiled for 5 minutes, run on an 8% acrylamide gel for 60 minutes at 125 V, and transferred to a nitrocellulose membrane. Membrane was bisected between 75kDa and 100kDa. Membranes were blocked with 3% bovine serum albumin in TBS-T for 1 hour at room temperature, and the upper membrane was immunostained with anti-DNMT3A (Abcam, 1:1000, ab13888) while the lower membrane was immunostained with anti-α-Tubulin (Abcam, 1:1000, ab52866) for 12-16 hours at 4°C. All primary and secondary antibodies were diluted in 3% Bovine Serum Albumin in TBS-T. Membranes were next washed with TBS-T, and then incubated for 1 hour at room temperature with IR-dye secondary antibody (IRDye 800CW Goat anti-Rabbit, or IRDye 800CW Goat anti-Mouse, LI-COR Biosciences, 1:15,000, product numbers: 926-32211 and 926-32210 respectively). Blots were then washed with PBS, imaged using the LiCOR Odyssey XCL system, and quantified using Image Studio Lite software (LI-COR Biosciences). DNMT3A and α-Tubulin levels were normalized to a standard curve, and protein levels are expressed as normalized DNMT3A values divided by normalized α-Tubulin values to enable comparison of DNMT3A levels between blots. For brain region analysis, sample sizes of n=4 per genotype (2 male and 2 female pairs) were used. For time course analysis, sample sizes of n=2 per genotype (1 male and 1 female pairs) were used for all time points except the 2-week timepoint in which n=6 (3 male and 3 female pairs) was used. Significance was assessed using a Student's T-Test for brain regions, and a two-way ANOVA considering genotype and time to determine if there was a detectable difference in protein expression over development.

Bisulfite sequencing

Whole genome bisulfite sequencing from cortical cultures. Samples were chosen for whole genome bisulfite sequencing if mutant and WT samples expressed equal amounts of *Dnmt3a* mRNA as measured by qRT-PCR (Figure 3.8F). DNA from cortical cultures was bisulfite converted and prepared for sequencing using the Tecan Ovation Ultralow Methyl-Seq Kit (Tecan, 0335-32) and the Epitect Bisulfite Kit (Qiagen, 59824) was used for bisulfite conversion. We used alternate bisulfite conversion cycling conditions ([95°C, 5 min; 60°C, 20 min] x 4 cycles, 20°C hold) to ensure lowest possible bisulfite non-conversion rate. Libraries were PCR-amplified for 10-11 cycles. Libraries were then pooled and sequenced at a depth of 0.01-0.03x genomic coverage using an Illumina MiSeq 2x150 through the Spike-In Cooperative at Washington University in St. Louis. Significance was assessed using a one-sample student's t-test, as we are comparing groups normalized to WT and GFP back to the normalized value of 1.

Whole genome bisulfite sequencing from tissue. DNA was isolated from tissue using the DNEasy Kit (Qiagen). 300 ng of DNA was prepared for sequencing using the Ovation Ultralow

Methyl-Seq Kit (Tecan, 0335-32) with and the Epitect Bisulfite Kit (Qiagen, 59824) was used for bisulfite conversion. For these samples, 300 ng of DNA was fragmented for 45 seconds with the Covaris E220 sonicator (10% Duty Factory, 175 Peak Incidence Power, 200 cycles per burst, milliTUBE 200µL AFA Fiber). DNA was then purified using 0.7 volumes of Agencourt Beads to select for long DNA inserts for sequencing. We used alternate bisulfite conversion cycling conditions ([95°C, 5 min; 60°C, 20 min] x 4 cycles, 20°C hold) to ensure lowest possible bisulfite non-conversion rate. Libraries were PCR-amplified for 12 cycles. Libraries were then pooled and sequenced using an Illumina MiSeq 2x150 through the Spike-In Cooperative at Washington University in St. Louis. Samples for shallow-depth sequencing (Figure 3.4A,B) were sequenced at 0.01-0.03x genomic coverage. For brain region and liver methylation, n=2 per genotype per region (one male pair, one female pair). For developmental time course methylation, n=3-4 per genotype per timepoint, with at least one male and one female pair. 8-week cortex samples for deep sequencing (Figure 3.5, Figure 3.6A-D, Figure 3.13D-H) were sequenced at 25-27x coverage per genotype using two male and two female WT-DNMT3A$^{KO/+}$ pairs (n=4 per genotype). Samples were sequenced either using the HiSeq 3000 or NextSeq 500 at 2x150 or 1x150, respectively. For shallow sequencing experiments, significance was assessed using either a two-sample Student's T-Test to compare global methylation values of the cortex between two genotypes or using a two-way ANOVA to compare broad methylation changes across a variety of brain regions or timepoints. Genomic element comparisons were done using two-sample Student's T-Tests with Bonferroni correction.

Oxidative bisulfite sequencing from tissue. DNA was isolated from tissue using the DNEasy Kit (Qiagen, 69504). 450 ng of DNA was prepared for sequencing using the Ovation Ultralow Methyl-Seq Kit (Tecan, 0335-32) with TrueMethyl oxBS plugin (Tecan, 0414-32). For these samples, 450 ng of DNA was fragmented for 45 seconds with the Covaris E220 sonicator (10% Duty Factory, 175 Peak Incidence Power, 200 cycles per burst, milliTUBE 200µL AFA Fiber). DNA was then purified using 0.7 volumes of Agencourt Beads to select for long DNA inserts for sequencing. 2/3 of the sample (~300 ng of DNA) was used for OxBS libraries, whereas the remaining 1/3 (~150 ng of DNA) was used for bisulfite libraries. We used alternate bisulfite conversion cycling conditions ([95°C, 5 min; 60°C, 20 min] x 2 cycles; 95°C, 5 min; 60°C, 40 min; 95°C, 5 min; 60°C, 45 min; 20°C hold) to ensure lowest possible bisulfite non-conversion rate. Bisulfite and oxidative bisulfite libraries were PCR-amplified for 11 and 13 cycles respectively. Libraries were then pooled and sequenced using an Illumina MiSeq 2x150 through the Spike-In Cooperative at Washington University in St. Louis. Samples were sequenced at 0.8-2.2x genomic coverage per replicate (two male replicates per genotype). Genomic element comparisons were done using two-sample Student's T-Tests with Bonferroni correction.

Whole-genome bisulfite analysis

Bisulfite sequencing analysis was performed as previously described (Clemens et al., 2019). Briefly, data were adapter-trimmed, mapped to mm9, then deduplicated and called for methylation using BS-seeker2. Methylation levels across regions were assessed using bedtools map -o sum, summing the number of reads mapping to Cs (supporting mC) and the amount of coverage in the region, then dividing those two numbers (Quinlan and Hall, 2010). Hydroxymethylation was calculated as the percent methylation found in the BS-seq minus the percent methylation found in the matching oxBS-seq. Due to count noise, this occasionally resulted in apparent negative hydroxymethylation. During bisulfite sequencing not all DNA can be efficiently bisulfite converted. Though our methods should maximize the amount of converted unmethylated C, there is still a small percentage of unmethylated cytosines that are called as methylated due to non-conversion (0.2-0.3%). Due to this non-conversion, very lowly methylated regions (e.g. mCA at CpG islands) may not show the same percent reduction in mCA as highly methylated regions. Data were visualized using the UCSC genome browser (http://genome.ucsc.edu) (Kent et al., 2002). CpG islands were obtained from the UCSC table browser (Haeussler et al., 2019), and CpG Shores were defined as the 8kb surrounding them. Average methylation per-sample is normally distributed in all regions examined, and variance between genotypes is similar, fitting the assumptions of a 2-sample t-test. Methylation levels for individual elements are not necessarily normally distributed, so non-parametric tests were used instead. This approach to mirrors multiple other studies quantifying mCA levels across the genome (Lister et al., 2013; Luo et al., 2017) where *de novo* calling of basepair resolution mCA DMRs is avoided due to low statistical power because of low mCA/CA levels at any individual site, and instead focuses on quantification of mCA/CA levels in populations of pre-defined regions of known functional importance.

Differentially methylated region detection

BSmooth (Hansen et al., 2012) was used to call differentially CpG methylated regions between DNMT3A$^{KO/+}$ and WT mice, using four biological replicates. CG sites were filtered for requiring at least 2x coverage in all replicates and differentially methylated regions were called with a statistical threshold of t-stat >2.0. These regions were further filtered for a length >100 bp and a requirement that the smoothed per-rep methylation values were consistent. For hypomethylated regions all WT mCG/CG values needed to be greater than any KO mCG/CG value, and for hypermethylated regions all KO methylation values needed to be higher than all WT methylation values. Data fit the assumptions and requirements of BSmooth. Data were distributed evenly between chromosomes, and the overlap between DMRs and regions of interest fit a hypergeometric distribution, making a fisher's exact test appropriate. In order to calculate an expected overlap of DMRs and genomic regions, for each chromosome, an equal number of resampled DMRs were placed, at random throughout the chromosome, as true DMRs, ensuring

that the two do not overlap. From this, the overlap of genomic regions with resampled DMRs was compared to the same overlap with true DMRs.

RNA sequencing

Total RNA isolation was carried out as previously described (Clemens et al., 2019). In brief, cerebral cortex was dissected in ice-cold PBS from DNMT3A$^{KO/+}$ and WT littermates at 8 weeks of age (n=7 pairs, 3 male, 4 female). Cortex was lysed in RLT buffer following the RNeasy Mini Kit (Qiagen, 74104). RNA libraries were generated from 250 ng of RNA with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB) using a modified amplification protocol (37°C, 15 minutes; 98°C, 30 seconds; [98°C, 10 seconds; 65°C, 30 seconds; 72°C, 30 seconds]x13; 72°C, 5 minutes; 4°C hold). RNA libraries were sequenced using Illumina HiSeq3000 1x50bp with the Genome Technology Access Center at Washington University in St. Louis, typically yielding 15-30 million single-end reads per sample.

RNA sequencing analysis

RNA sequencing analysis was performed as previously described (Clemens et al., 2019). Briefly, raw FASTQ files were trimmed with Trim Galore and rRNA sequences were filtered out with Bowtie. Remaining reads were aligned to mm9 using STAR (Dobin et al., 2013) with the default parameters. Reads mapping to multiple regions in the genome were then filtered out, and uniquely mapping reads were converted to BED files and separated into intronic and exonic reads. Finally, reads were assigned to genes using bedtools coverage -counts (Quinlan and Hall, 2010).

For gene annotation we defined a "flattened" list of longest transcript forms for each gene, generated on Ensgene annotations and obtained from the UCSC table browser. For each gene, Ensembl IDs were matched up to MGI gene names. Then, for each unique MGI gene name, the most upstream Ensgene TSS and the most downstream TES were taken as that gene's start and stop. Based on these Ensembl gene models, we defined TSS regions and gene bodies. Differentially expressed genes were identified using a Wald test through DESeq2, running using default parameters on exonic reads from the DNMT3A$^{KO/+}$ and WT.

To assess effects on other noncoding RNAs in the DNMT3A$^{KO/+}$ strain, we realigned our RNA-seq data to all mouse noncoding RNAs from RNAcentral (The RNAcentral Consortium 2019, *Nucleic Acids Res.*), then ran DESeq2 on the counts from each noncoding transcript.

Chromatin immunoprecipitation protocol

Chromatin immunoprecipitation was performed as previously described (Clemens et al., 2019; Cohen et al., 2011). Cerebral cortex was dissected on ice in PBS from DNMT3A$^{KO/+}$ and WT

littermates at 8-weeks old (n=5 pairs, 3 male, 2 female). The tissue was flash-frozen in liquid nitrogen and stored at -80°C. Chromatin was fragmented with the Covaris E220 sonicator (5% Duty Factory, 140 Peak Incidence Power, 200 cycles per burst, milliTUBE 1mL AFA Fiber). ChIP was performed with H3K27ac antibody (0.025-0.1µg; Abcam, ab4729) and libraries were generated using Ovation Ultralow Library System V2 (Tecan, 0344NB-32). Libraries were sequenced using Illumina HiSeq 3000 with the Genome Technology Access Center at Washington University in St. Louis, typically yielding 15-40 million single-end reads per sample.

Chromatin immunoprecipitation analysis

ChIP sequencing analysis was performed as previously described (Clemens et al., 2019). Briefly, reads were mapped to mm9 using bowtie2 and reads were extended based on library sizes and deduplicated. Bedtools coverage –counts was used to quantify ChIP signal at the transcriptional start site (TSS), gene body (GB), and transcriptional end site (TES) (Quinlan and Hall, 2010). edgeR was then used to determine differential ChIP-signal across genotypes. Data were visualized using the UCSC genome browser (http://genome.ucsc.edu) (Kent et al., 2002).

Controlled resampling

A similar resampling approach was used as previously described (Clemens et al., 2019). Briefly, for every entry in a sample set (e.g. DNMT3A-dysregulated genes), an entry in the control set (e.g. all other genes) with a similar desired characteristic (e.g. expression) was selected, generating a control set of the same size and variable distribution as the sample set.

Identification of dysregulated enhancers

Enhancer regions from Clemens et al. 2019 were used, and enhancers dysregulated in the $DNMT3A^{KO/+}$ were called using the same method. Briefly, H3K27ac ChIP-seq reads were quantified in all acetyl peak regions, and edgeR was used to identify peaks with significantly different amounts of H3K27ac signal. Peak regions were then divided into promoters, enhancers, and non-identified peaks. Data fits the assumptions of BSmooth. Overlap between misregulated enhancers in different genotypes fit a hypergeometric distribution.

GAGE

Gene set enrichment analysis for the gene sets described was performed using the Generally Applicable Gene-set Enrichment (GAGE) program (Luo et al., 2009). The NDD models for comparison were chosen by searching for gene expression datasets meeting the following criteria: 1) NDD/ASD models that have at least some similar features to DNMT3A disorders (e.g. ID, ASD) 2) generated with the RNA-seq approach 3) analyzed brain tissue with enough similarity to our cortical analysis to justify a reasonable comparison. Analysis was performed

directionally on the shrunken, log-normalized exonic fold changes from DESeq2 analysis of DNMT3A$^{KO/+}$ versus WT RNA-seq data. For each gene set, fold changes of genes in that set were compared to a background of all expressed genes. Gene sets with an FDR q-value below 0.1 and an adjusted p-value below 0.5 following expression matched resampling repeated 1,000 times were considered statistically significant. Gene sets were selected for analysis from both human and mouse studies of autism associated genes. SFARI genes (Abrahams et al., 2013) with scores of equal to or less than 3 were considered. Date accessed: 6/20/2019.

GSEA

Gene Set Enrichment Analysis (GSEA) (version 7.0, the Broad Institute of MIT and Harvard, http://software.broadinstitute.org/gsea/downloads.jsp) was performed on shrunken, log-normalized exonic fold changes from DESeq2 between DNMT3A^{KO+} and WT RNA-seq data. GSEA calculated a gene set Enrichment Score (ES) that analyzed genes were enriched in the biological signal conduction on the MsigDB (Molecular Signatures Database, http://software.broadinstitute.org/gsea/msigdb). Background was set to all expressed genes in this study and 1,000 permutations were set to generate a null distribution for enrichment score in the hallmark gene sets and functional annotation gene sets. The gene sets database used for enrichment analysis were 'c5.all.v7.0.symbols.gmt', 'c5.bp.v7.0.symbols.gmt', 'c5.cc.v7.0.symbols.gmt'and 'c5.mf.v7.0.symbols.gmt' and FDR <0.1 was defined as the cut-off criteria for significance.

Craniofacial morphological analyses

A total of 24 sex-matched littermate paired mice (DNMT3A$^{KO/+}$ n=12, 7 male, 5 female; WT n=12, 7 male, 5 female) across 3 time-points (8 weeks DNMT3A$^{KO/+}$ n=4, WT n=4; 20 weeks DNMT3A$^{KO/+}$ n=4, WT n=4; 25 weeks DNMT3A$^{KO/+}$ n=4, WT n=4) were fixed in 4% paraformaldehyde through intracardiac perfusion. Whole mouse heads were scanned at the Musculoskeletal Research Center at Washington University in St. Louis using a Scanco μCT40 machine. CT images had voxel dimensions of 0.018 millimeters and were reconstructed on a 2048x2048 pixel grid. The CT images were converted to 8bit images using ImageJ (https://imagej.nih.gov/ij/) and surface reconstructions were acquired in Avizo (http://www.vsg3d.com/). Thirty-five three-dimensional landmarks were collected from surface reconstructions of the cranium and mandible using Stratovan Checkpoint (https://www.stratovan.com/products/checkpoint).

Generalized Procrustes Analysis in MorphoJ software was used to explore the differences and similarities of shape between the DNMT3A$^{KO/+}$ mice and their WT littermates as previously described (Hill et al., 2013). To control for possible differences in size, the landmark coordinate data were natural log-transformed and analyzed with a linear regression model. Additionally, to

localize differences in form to specific linear distances, landmark data were analyzed using Euclidean Distance Matrix Analysis (EDMA).

Bone length measurements

We chose to quantify long bones that may directly relate to the height phenotype seen in patients. A total of 24 sex-matched littermate paired mice ($DNMT3A^{KO/+}$ n=12, 7 male, 5 female; WT n=12, 7 male, 5 female) across 3 time-points (8 weeks $DNMT3A^{KO/+}$ n=4, WT n=4; 20 weeks $DNMT3A^{KO/+}$ n=4, WT n=4; 25 weeks $DNMT3A^{KO/+}$ n=4, WT n=4) were fixed in 4% paraformaldehyde through intracardiac perfusion. Decapitated mouse bodies were scanned at the Musculoskeletal Research Center at Washington University in St. Louis using a Faxitron Model UltraFocus100 Dual X-Ray machine. Bone lengths were measured using ImageJ. Data were taken over three age time-points: 8 weeks, 20 weeks, and 25 weeks of age for male and female mice. There was no significant difference in bone lengths based upon sex, but there was a difference based by age. To normalize for this age effect, data were expressed as $DNMT3A^{KO/+}$ bone lengths normalized to the WT lengths within groups. This also mirrors how human data are presented. Left and right bones were measured and the larger was used for analysis.

Experimental design

Authenticated cell lines from ATCC (HEK293T, NEURO-2A) were used, and no mycoplasma contamination testing was needed. Sample sizes were chosen based upon previously published studies using similar techniques. Statistical tests and exclusion criteria (values beyond 2 standard deviations of the group mean) were similar to that of previously published studies and indicated in the appropriate methods. For all animal experiments, experimenters were blinded to genotype during data collection. No treatment conditions were used, so no samples or animals were allocated to experimental groups and no randomization was needed. Tests that assume equal variance were only run if group variances were similar, otherwise alternative tests were used.

3.6 Acknowledgements

We thank the Division of Comparative Medicine at Washington University in Saint Louis for their assistance with mouse husbandry and veterinary support. We are grateful for sequencing from the Genome Technology Access Center and extensive sequencing support from J. Hoisington-Lopez and M. Crosby at the Center for Genome Sciences at Washington University in Saint Louis. We thank members of the Dougherty lab including S. Maloney, K. McCullough, and M. Rieger for assistance in USV experiments. We thank J. Edwards, J. Goodman, and J. Yi for critical feedback on the experimental design and manuscript. We thank A. Smith and T. Ley for helpful discussions. This publication was made possible in part by Grant Number UL1 RR024992 from the NIH-National Center for Research Resources (NCRR). This work was also supported by NIH 5T32GM00815133 and F31HD1000098 to D.L.C, by NIH 5T32GM007067

and F31NS108574 to A.W.C., and by the Klingenstein-Simons Fellowship Fund, the G. Harold and Leila Y. Mathers Foundation, the Brain and Behavior Research Foundation, the Simons Foundation for Autism Research Initiative, and NIMH R01MH117405 to H.W.G.

Declaration of interests

The authors declare no competing interests.

3.7 References cited

1. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism *4*, 36.
2. Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2 , encoding methyl-CpG-binding protein 2. Nat. Genet. *23*, 185–188.
3. Anderson, G.R., Maxeiner, S., Sando, R., Tsetsenis, T., Malenka, R.C., and Südhof, T.C. (2017). Postsynaptic adhesion GPCR latrophilin-2 mediates target recognition in entorhinal-hippocampal synapse assembly. J. Cell Biol. *216*, 3831–3846.
4. Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. Dis. Model. Mech. *3*, 133–135.
5. Barnes, T.D., Rieger, M.A., Dougherty, J.D., and Holy, T.E. (2017). Group and Individual Variability in Mouse Pup Isolation Calls Recorded on the Same Day Show Stability. Front. Behav. Neurosci. *11*.
6. Boxer, L.D., Renthal, W., Greben, A.W., Whitwam, T., Silberfeld, A., Stroud, H., Li, E., Yang, M.G., Kinde, B., Griffith, E.C., et al. (2019). MeCP2 Represses the Rate of Transcriptional Initiation of Highly Methylated Long Genes. Mol. Cell.
7. Boyle, M.P. (2006). Forebrain Glucocorticoid Receptors Modulate Anxiety-Associated Locomotor Activation and Adrenal Responsiveness. J. Neurosci. *26*, 1971–1978.
8. Chang, Y., Cole, T.B., and Costa, L.G. (2017). Behavioral Phenotyping for Autism Spectrum Disorders in Mice. Curr. Protoc. Toxicol. *72*.
9. Chen, L., Chen, K., Lavery, L.A., Baker, S.A., Shaw, C.A., Li, W., and Zoghbi, H.Y. (2015). MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. Proc. Natl. Acad. Sci. *112*, 5509–5514.
10. Cholewa-Waclaw, J., Bird, A., von Schimmelmann, M., Schaefer, A., Yu, H., Song, H., Madabhushi, R., and Tsai, L.-H. (2016). The Role of Epigenetic Mechanisms in the Regulation of Gene Expression in the Nervous System. J. Neurosci. *36*, 11427–11434.
11. Clemens, A.W., Wu, D.Y., Moore, J.R., Christian, D.L., Zhao, G., and Gabel, H.W. (2019). MeCP2 Represses Enhancers through Chromosome Topology-Associated DNA Methylation. Mol. Cell.

12. Cohen, S., Gabel, H.W., Hemberg, M., Hutchinson, A.N., Sadacca, L.A., Ebert, D.H., Harmin, D.A., Greenberg, R.S., Verdine, V.K., Zhou, Z., et al. (2011). Genome-Wide Activity-Dependent MeCP2 Phosphorylation Regulates Nervous System Development and Function. Neuron *72*, 72–85.
13. Cole, C.B., Russler-Germain, D.A., Ketkar, S., Verdoni, A.M., Smith, A.M., Bangert, C.V., Helton, N.M., Guo, M., Klco, J.M., O'Laughlin, S., et al. (2017). Haploinsufficiency for DNA methyltransferase 3A predisposes hematopoietic cells to myeloid malignancies. J. Clin. Invest. *127*, 3657–3674.
14. Das, D., Zalewski, J.K., Mohan, S., Plageman, T.F., VanDemark, A.P., and Hildebrand, J.D. (2014). The interaction between Shroom3 and Rho-kinase is required for neural tube morphogenesis in mice. Biol. Open *3*, 850–860.
15. Deshwar, A.R., Martin, N., Shannon, P., and Chitayat, D. (2020). A homozygous pathogenic variant in *SHROOM3* associated with anencephaly and cleft lip and palate. Clin. Genet. *98*, 299–302.
16. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
17. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
18. Dougherty, J.D., Maloney, S.E., Wozniak, D.F., Rieger, M.A., Sonnenblick, L., Coppola, G., Mahieu, N.G., Zhang, J., Cai, J., Patti, G.J., et al. (2013). The Disruption of *Celf6*, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. J. Neurosci. *33*, 2732 LP – 2753.
19. Duyar, I., and Pelin, C. (2003). Body height estimation based on tibia length in different stature groups. Am. J. Phys. Anthropol. *122*, 23–27.
20. Fazel Darbandi, S., Robinson Schwartz, S.E., Qi, Q., Catta-Preta, R., Pai, E.L.-L., Mandell, J.D., Everitt, A., Rubin, A., Krasnoff, R.A., Katzman, S., et al. (2018). Neonatal Tbr1 Dosage Controls Cortical Layer 6 Connectivity. Neuron *100*, 831-845.e7.
21. Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T.N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L.G., Muzny, D.M., et al. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. Npj Genomic Med. *4*, 1–14.
22. Fox, J., and Weisberg, S. (2011). An R Companion to Applied Regression (SAGE Publications).
23. 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 *522*, 89–93.
24. Gallitano-Mendel, A., Wozniak, D.F., Pehek, E.A., and Milbrandt, J. (2008). Mice Lacking the Immediate Early Gene Egr3 Respond to the Anti-Aggressive Effects of Clozapine Yet are Relatively Resistant to its Sedating Effects. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. *33*, 1266–1275.

25. Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*, eaat8127.
26. Gompers, A.L., Su-Feher, L., Ellegood, J., Copping, N.A., Riyadh, M.A., Stradleigh, T.W., Pride, M.C., Schaffler, M.D., Wade, A.A., Catta-Preta, R., et al. (2017). Germline Chd8 haploinsufficiency alters brain development in mouse. Nat. Neurosci. *20*, 1062–1073.
27. Gowher, H., and Jeltsch, A. (2018). Mammalian DNA methyltransferases: new discoveries and open questions. Biochem. Soc. Trans. *46*, 1191–1202.
28. Grady, R.M., Wozniak, D.F., Ohlemiller, K.K., and Sanes, J.R. (2006). Cerebellar Synaptic Defects and Abnormal Motor Behavior in Mice Lacking α- and β-Dystrobrevin. J. Neurosci. *26*, 2841–2851.
29. Guilmatre, A., Huguet, G., Delorme, R., Bourgeron, T. (2014) The emerging role of SHANK genes in neuropsychiatric disorders. Dev. Neurobiol. *74*, 113-122.
30. 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.
31. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.-Y., and Pellegrini, M. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics *14*, 774.
32. Guo, X., Wang, L., Li, J., Ding, Z., Xiao, J., Yin, X., He, S., Shi, P., Dong, L., Li, G., et al. (2015). Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. Nature *517*, 640–644.
33. Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47*, D853–D858.
34. Hansen, K.D., Langmead, B., and Irizarry, R.A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. *13*, R83.
35. Hartman, R.E., Wozniak, D.F., Nardi, A., Olney, J.W., Sartorius, L., and Holtzman, D.M. (2001). Behavioral Phenotyping of GFAP-ApoE3 and -ApoE4 Transgenic Mice: ApoE4 Mice Show Profound Working Memory Impairments in the Absence of Alzheimer's-like Neuropathology. Exp. Neurol. *170*, 326–344.
36. Hill, C.A., Martínez-Abadías, N., Motch, S.M., Austin, J.R., Wang, Y., Jabs, E.W., Richtsmeier, J.T., and Aldridge, K. (2013). Postnatal brain and skull growth in an Apert syndrome mouse model. Am. J. Med. Genet. A. *161*, 745–757.
37. Holy, T.E., and Guo, Z. (2005). Ultrasonic Songs of Male Mice. PLoS Biol. *3*, e386.
38. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
39. Ip, J.P.K., Mellios, N., and Sur, M. (2018). Rett syndrome: insights into genetic, molecular and circuit mechanisms. Nat. Rev. Neurosci. *19*, 368–382.
40. Jeltsch, A., Broche, J., and Bashtrykov, P. (2018). Molecular Processes Connecting DNA Methylation Patterns with DNA Methyltransferases and Histone Modifications in Mammalian Genomes. Genes *9*, 566.

41. Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.
42. Katayama, Y., Nishiyama, M., Shoji, H., Ohkawa, Y., Kawamura, A., Sato, T., Suyama, M., Takumi, T., Miyakawa, T., and Nakayama, K.I. (2016). CHD8 haploinsufficiency results in autistic-like phenotypes in mice. Nature *537*, 675–679.
43. Kazdoba, T.M., Leach, P.T., Silverman, J.L., and Crawley, J.N. (2014). Modeling fragile X syndrome in the Fmr1 knockout mouse. Intractable Rare Dis. Res. *3*, 118–133.
44. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.
45. Kinde, B., Gabel, H.W., Gilbert, C.S., Griffith, E.C., and Greenberg, M.E. (2015). Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. Proc. Natl. Acad. Sci. *112*, 6800–6806.
46. Kinde, B., Wu, D.Y., Greenberg, M.E., and Gabel, H.W. (2016). DNA methylation in the gene body influences MeCP2-mediated gene repression. Proc. Natl. Acad. Sci. *113*, 15114–15119.
47. Lagger, S., Connelly, J.C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B.H., Yu, M., He, C., Sanguinetti, G., Sowers, L.C., et al. (2017). MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. PLOS Genet. *13*, e1006793.
48. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
49. Lazic, S.E. (2015). Analytical strategies for the marble burying test: avoiding impossible predictions and invalid p-values. BMC Res. Notes *8*, 141.
50. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
51. 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*.
52. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.
53. Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.
54. Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., and Woolf, P.J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics *10*, 161.
55. Lyst, M.J., and Bird, A. (2015). Rett syndrome: a complex disorder with simple roots. Nat. Rev. Genet. *16*, 261–275.
56. Maloney, S.E., Rieger, M.A., Al-Hasani, R., Bruchas, M.R., Wozniak, D.F., and Dougherty, J.D. (2019a). Loss of CELF6 RNA binding protein impairs cocaine conditioned place preference and contextual fear conditioning. Genes Brain Behav. *18*, e12593.

57. Maloney, S.E., Yuede, C.M., Creeley, C.E., Williams, S.L., Huffman, J.N., Taylor, G.T., Noguchi, K.N., and Wozniak, D.F. (2019b). Repeated neonatal isoflurane exposures in the mouse induce apoptotic degenerative changes in the brain and relatively mild long-term behavioral deficits. Sci. Rep. *9*, 1–14.
58. Manno, R., Witte, J., and Papouin, T. (2020). A modular set-up to run a large line of behavioral testing in mice in a single space. Current Protocols in Neuroscience *93, e102*.
59. 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.
60. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.
61. Mei, Y., Monteiro, P., Zhou, Y., Kim, J.-A., Gao, X., Fu, Z., and Feng, G. (2016). Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. Nature *530*, 481–484.
62. Moy, S.S., Nadler, J.J., Perez, A., Barbaro, R.P., Johns, J.M., Magnuson, T.R., Piven, J., and Crawley, J.N. (2004). Sociability and preference for social novelty in five inbred strains: an approach to assess autistic-like behavior in mice. Genes Brain Behav. *3*, 287–302.
63. Nguyen, S., Meletis, K., Fu, D., Jhaveri, S., and Jaenisch, R. (2007). Ablation of de novo DNA methyltransferase Dnmt3a in the nervous system leads to neuromuscular defects and shortened lifespan. Dev. Dyn. *236*, 1663–1676.
64. Nord, A.S., and West, A.E. (2019). Neurobiological functions of transcriptional enhancers. Nat. Neurosci.
65. Nothjunge, S., Nührenberg, T.G., Grüning, B.A., Doppler, S.A., Preissl, S., Schwaderer, M., Rommel, C., Krane, M., Hein, L., and Gilsbach, R. (2017). DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. Nat. Commun. *8*, 1667.
66. Nygaard, K.R., Maloney, S.E., and Dougherty, J.D. (2019). Erroneous inference based on a lack of preference within one group: Autism, mice, and the social approach task. Autism Res. *12*, 1171–1183.
67. Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.
68. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell *155*, 1008–1021.
69. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
70. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
71. Russler-Germain, D.A., Spencer, D.H., Young, M.A., Lamprecht, T.L., Miller, C.A., Fulton, R., Meyer, M.R., Erdmann-Gilmore, P., Townsend, R.R., Wilson, R.K., et al.

(2014). The R882H DNMT3A Mutation Associated with AML Dominantly Inhibits Wild-Type DNMT3A by Blocking Its Ability to Form Active Tetramers. Cancer Cell *25*, 442–454.
72. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233.
73. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. BioRxiv 484113.
74. Sawyer, S.L., Hartley, T., Dyment, D.A., Beaulieu, C.L., Schwartzentruber, J., Smith, A., Bedford, H.M., Bernard, G., Bernier, F.P., Brais, B., et al. (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care: Whole-exome sequencing for rare disease diagnosis. Clin. Genet. *89*, 275–284.
75. Sceniak, M.P., Lang, M., Enomoto, A.C., James Howell, C., Hermes, D.J., and Katz, D.M. (2016). Mechanisms of Functional Hypoconnectivity in the Medial Prefrontal Cortex of Mecp2 Null Mice. Cereb. Cortex N. Y. NY *26*, 1938–1956.
76. Sendžikaitė, G., Hanna, C.W., Stewart-Morgan, K.R., Ivanova, E., and Kelsey, G. (2019). A DNMT3A PWWP mutation leads to methylation of bivalent chromatin and growth retardation in mice. Nat. Commun. *10*, 1–16.
77. Sessa, A., Fagnocchi, L., Mastrototaro, G., Massimino, L., Zaghi, M., Indrigo, M., Cattaneo, S., Martini, D., Gabellini, C., Pucci, C., et al. (2019). SETD5 Regulates Chromatin Methylation State and Preserves Global Transcriptional Fidelity during Brain Development and Neuronal Wiring. Neuron *104*, 271-289.e13.
78. Silverman, J.L., Turner, S.M., Barkan, C.L., Tolu, S.S., Saxena, R., Hung, A.Y., Sheng, M., and Crawley, J.N. (2011). Sociability and motor functions in Shank1 mutant mice. Brain Res. *1380*, 120–137.
79. Simola, N., and Granon, S. (2019). Ultrasonic vocalizations as a tool in studying emotional states in rodent models of social behavior and brain disease. Neuropharmacology *159*, 107420.
80. Spencer, D.H., Russler-Germain, D.A., Ketkar, S., Helton, N.M., Lamprecht, T.L., Fulton, R.S., Fronick, C.C., O'Laughlin, M., Heath, S.E., Shinawi, M., et al. (2017). CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. Cell *168*, 801-816.e13.
81. Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.
82. Stroud, H., Su, S.C., Hrvatin, S., Greben, A.W., Renthal, W., Boxer, L.D., Nagy, M.A., Hochbaum, D.R., Kinde, B., Gabel, H.W., et al. (2017). Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. Cell *171*, 1151-1164.e16.
83. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.
84. Takumi, T., Tamada, K., Hatanaka, F., Nakai, N., and Bolton, P.F. (2019). Behavioral neuroscience of autism. Neurosci. Biobehav. Rev. S0149763418303725.
85. Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., Del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nat. Genet. *46*, 385–388.
86. Tatton-Brown, K., Zachariou, A., Loveday, C., Renwick, A., Mahamdallie, S., Aksglaede, L., Baralle, D., Barge-Schaapveld, D., Blyth, M., Bouma, M., et al. (2018). The Tatton-Brown-Rahman Syndrome: A clinical study of 55 individuals with de novo constitutive DNMT3A variants. Wellcome Open Res. *3*, 1–16.
87. Tillotson, R., Selfridge, J., Koerner, M.V., Gadalla, K.K.E., Guy, J., De Sousa, D., Hector, R.D., Cobb, S.R., and Bird, A. (2017). Radically truncated MeCP2 rescues Rett syndrome-like neurological defects. Nature *550*, 398–401.
88. Tilot, A.K., Bebek, G., Niazi, F., Altemus, J.B., Romigh, T., Frazier, T.W., and Eng, C. (2016). Neural transcriptome of constitutional Pten dysfunction in mice and its relevance to human idiopathic autism spectrum disorder. Mol. Psychiatry *21*, 118–125.
89. de la Torre-Ubieta, L., Won, H., Stein, J.L., and Geschwind, D.H. (2016). Advancing the understanding of autism disease mechanisms through genetics. Nat. Med. *22*, 345–361.
90. Van Esch, H., Bauters, M., Ignatius, J., Jansen, M., Raynaud, M., Hollanders, K., Lugtenberg, D., Bienvenu, T., Jensen, L.R., Gécz, J., et al. (2005). Duplication of the MECP2 Region Is a Frequent Cause of Severe Mental Retardation and Progressive Neurological Symptoms in Males. Am. J. Hum. Genet. *77*, 442–453.
91. Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380–384.
92. Weinreb, C., and Raphael, B.J. (2016). Identification of hierarchical chromatin domains. Bioinformatics *32*, 1601–1609.
93. Wozniak, D.F., Hartman, R.E., Boyle, M.P., Vogt, S.K., Brooks, A.R., Tenkova, T., Young, C., Olney, J.W., and Muglia, L.J. (2004). Apoptotic neurodegeneration induced by ethanol in neonatal mice is associated with profound learning/memory deficits in juveniles followed by progressive functional recovery in adults. Neurobiol. Dis. *17*, 403–414.
94. 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.
95. Yang, M., Silverman, J.L., and Crawley, J.N. (2011). Automated Three-Chambered Social Approach Task for Mice. Curr. Protoc. Neurosci. *56*, 8.26.1-8.26.16.
96. Zhang, R.S., Liakath-Ali, K., and Südhof, T.C. (2020). Latrophilin-2 and latrophilin-3 are redundantly essential for parallel-fiber synapse function in cerebellum. ELife *9*, e54443.
97. 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.

98. Zhang, Z.M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S.B., Wang, Y., et al. (2018). Structural basis for DNMT3A-mediated de novo DNA methylation. Nature *554*, 387–391.

3.8 Figures

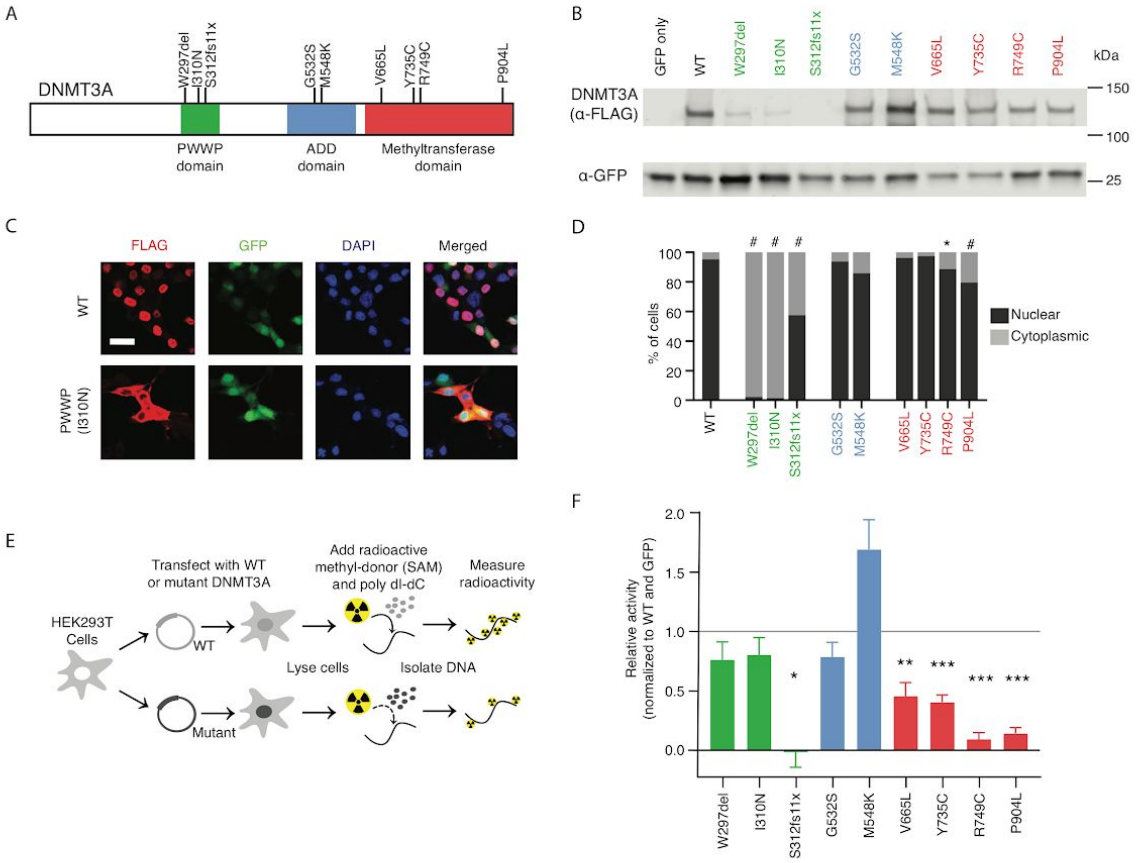

Figure 3.1. Disease-associated DNMT3A mutations disrupt distinct aspects of protein function.

A. Schematic of human DNMT3A protein showing canonical domains and disease-associated mutations (Sanders et al., 2015; Tatton-Brown et al., 2018) that were tested in this study.
B. Example immunoblot of DNMT3A mutant protein expression in Neuro-2a cells.
C. Example images of DNMT3A protein immunocytochemistry from wild type and PWWP domain mutant in Neuro-2a cells. Scale bar = 20μm.
D. Quantification of DNMT3A mutant protein localization (n=6-16 images; Generalized Linear Model test of percent nuclear expression per image for mutants compared to WT with Bonferroni correction).
E. Schematic of *in vitro* methylation assay for DNMT3A mutant proteins isolated from HEK293T cells.
F. Activity of DNMT3A mutant proteins in the *in vitro* methylation assay normalized to WT DNMT3A (set equal to 1) and GFP-only (set equal to 0) controls. (n=4-19; one-sample Student's T-Test with Bonferroni correction).

*, $P<0.05$; **, $P<0.01$; ***, $P<0.001$; #, $P<0.0001$. Bar graphs indicate mean and SEM.

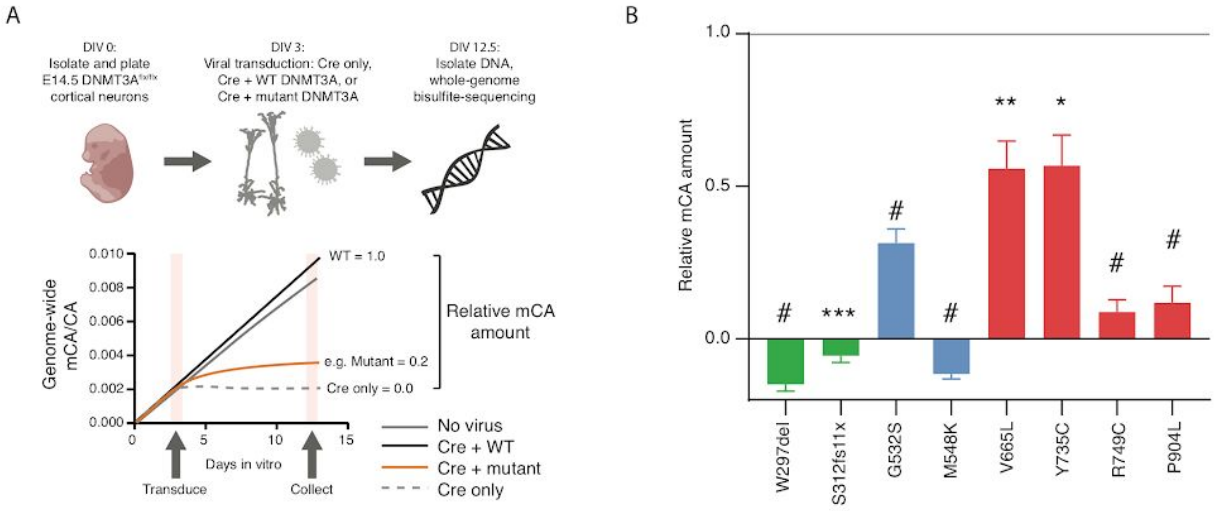

Figure 3.2. Disease-associated DNMT3A mutations prevent buildup of neuronal CA methylation.

A. Schematic of DNMT3A functional analysis in primary culture neurons. Cortical neurons are harvested from DNMT3A$^{flx/flx}$ mice at E14.5 and cultured. After 3 days *in vitro* (DIV), neurons are virally transduced with Cre recombinase and WT or mutant FLAG-tagged DNMT3A. On DIV 12.5, DNA and RNA are collected. Equal DNMT3A mRNA expression is verified by qRT-PCR (Figure S1F) and DNA is used for whole genome bisulfite sequencing analysis.

B. Relative mCA amount compared to Cre only (set equal to 0) and Cre+WT DNMT3A (set equal to 1) controls (n=4-11; one-sample Student's T-Test with Bonferroni correction).

*, P<0.05; **, P<0.01; ***, P<0.001; #, P<0.0001. Bar graphs indicate mean and SEM.

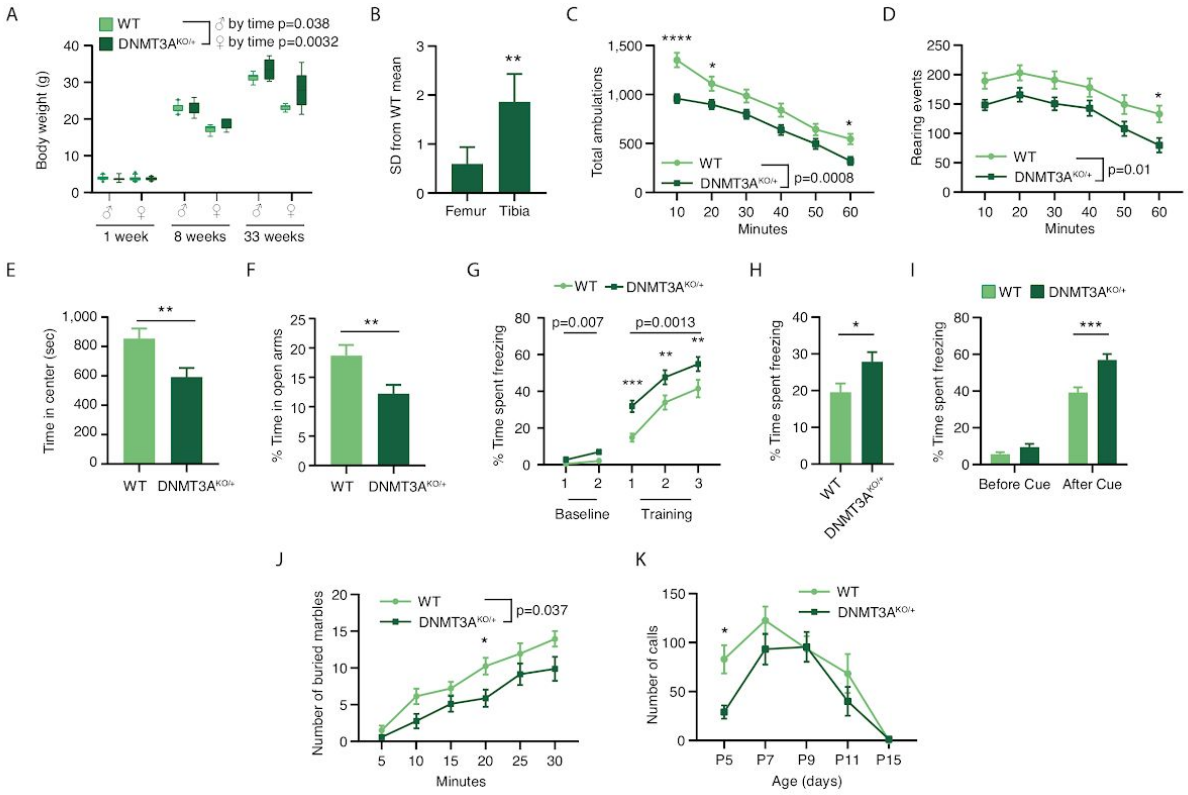

Figure 3.3. Heterozygous disruption of DNMT3A *in vivo* leads to growth and behavioral alterations.

A. Body weight of DNMT3A$^{KO/+}$ and WT mice at three developmental timepoints (Male P=0.038 genotype by age interaction effect, $F_{(2.50)}$=3.494, n=6-18; Female P=0.0032 genotype by age interaction effect, $F_{(2.48)}$=6.498; Female P=0.0016 genotype effect, $F_{(1.48)}$=11.18, n=5-17; two-way ANOVA).

B. Lengths of femur and tibia bones measured by dual X-ray imaging shown as standard deviations from the WT mean for the DNMT3A$^{KO/+}$ mice (n=12; paired Student's T-Test).

C. Total ambulations during 1-hour open-field testing, split into 10-minute bins (P=0.0008 effect by genotype, $F_{(1.46)}$=13.02, n=21,27; two-way repeated-measures ANOVA with Sidak's multiple comparison test).

D. Number of rearing events during 1-hour open-field testing, split into 10-minute bins (P=0.0103 effect by genotype, $F_{(1.46)}$=7.161, n=21,27; two-way repeated-measures ANOVA with Sidak's multiple comparison test).

E. Total time spent in the center zone of field during open-field testing (P=0.0075, n=21,27; unpaired Student's T-Test).

F. Percent of time mice spent in the open arms compared to all arms during first day of elevated-plus maze testing (P=0.0069; n=33,39; unpaired Student's T-Test).

G-I. Percent time spent freezing in

G. Conditioned fear training (Baseline: P=0.0071 effect by genotype, $F_{(1.50)}$=7.897; Cue: P=0.0013 effect by genotype, $F_{(1.50)}$=11.7; n=26; two-way repeated-measures ANOVA with Sidak's multiple comparisons test),
H. contextual fear trials (P=0.0215, n=26; unpaired Student's T-Test), and
I. cued fear trials (Before Cue: P=0.0606; After Cue: P=0.00014; n=26; unpaired Student's T-Test).
J. Quantification of marbles buried during 30 minutes of testing split into 5-minute bins (P=0.0374 effect by genotype, $F_{(1.25)}$=4.834, n=14,13; two-way repeated-measures ANOVA with Sidak's multiple comparison test).
K. Number of ultrasonic calls from pup isolated from the nest for 3-minute testing at indicated developmental time points; P=0.0378 effect by genotype, $F_{(1.285)}$=4.355, n=9-46; two-way ANOVA with Sidak's multiple comparisons test.

*, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001. Box plots contain 10^{th}-90^{th} percentiles of data, with remaining data represented as individual points. Line graphs and bar graphs indicate mean and SEM.

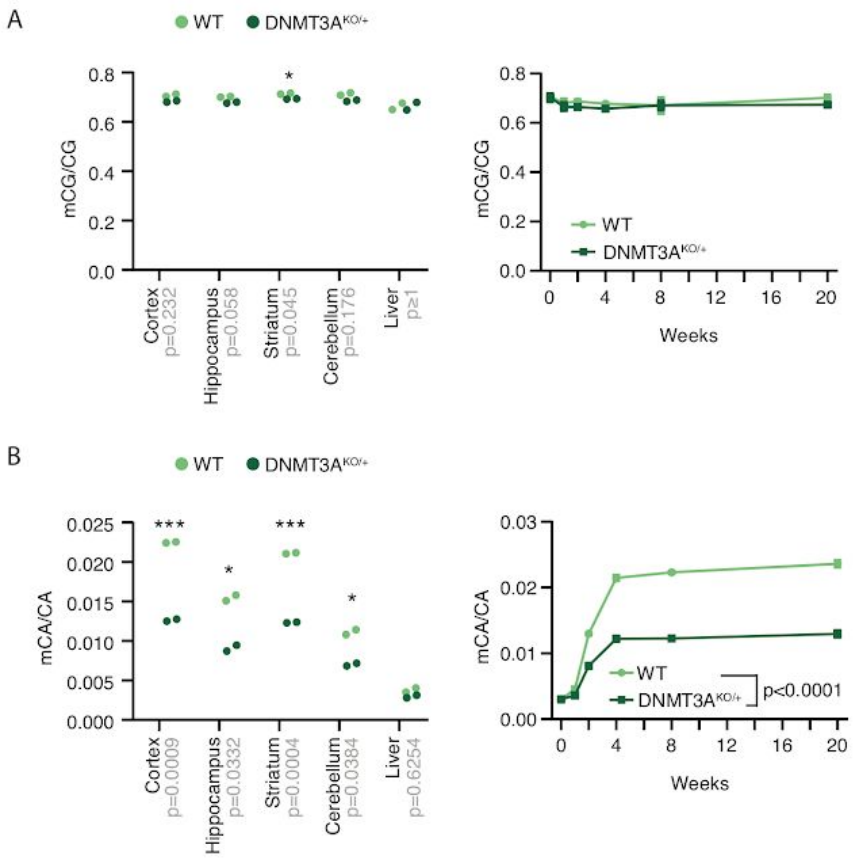

Figure 3.4. Global DNA methylation levels upon heterozygous loss of DNMT3A.

A. Global mCG levels in DNA isolated from brain regions of 8-week old mice (left) (unpaired Student's T-Test with Bonferroni correction), and developmental time course of global mCG in the cerebral cortex (right), as measured by sparse whole genome bisulfite sequencing (WGBS).

B. Global mCA levels in DNA isolated from brain regions of 8-week old mice (left) (unpaired Student's T-Test with Bonferroni correction), and developmental time course of global mCA in the cerebral cortex (right), as measured by sparse WGBS (P<0.0001 effect by genotype, $F_{(1,27)}$=1024; P<0.0001 effect by age $F_{(5,27)}$=884.6; n=3-4; two-way ANOVA).

Bonferroni corrected P-values indicated along the X-axis. *, P<0.05; ***, P<0.001. Line graphs indicate mean and SEM.

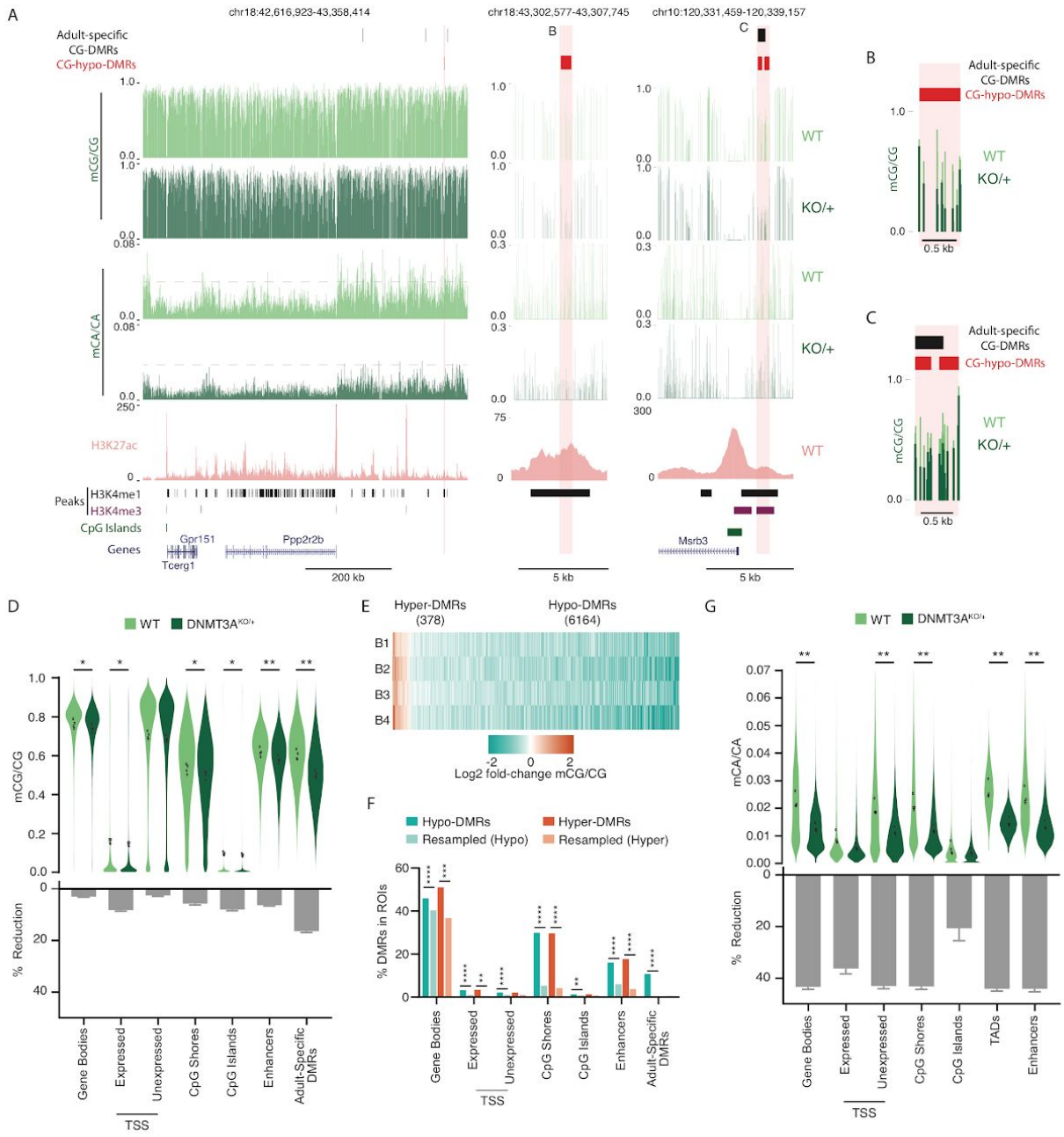

Figure 3.5. High-resolution analysis of DNA methylation changes in the DNMT3A$^{KO/+}$ cerebral cortex.

A. Genome browser views of mCA and mCG in WT and DNMT3A$^{KO/+}$ cerebral cortex as measured by high-depth WGBS. Broad view showing global reduction in mCA (left). Grey dashed line in mCA plots to facilitate visual comparison of global mCA levels between genotypes. Zoomed-in view of a DNMT3A$^{KO/+}$ CG-hypo-DMR that overlaps an

enhancer (center) and a DNMT3A$^{KO/+}$ CG-hypo-DMR at a CpG-island shore that overlaps with an adult-specific DMR (right). WT H3K27ac ChIP-seq signal (Clemens et al., 2019)**,** peaks of enhancer-associated H3K4me1 (Stamatoyannopoulos et al., 2012), peaks of promoter-associated H3K4me3 (Stamatoyannopoulos et al., 2012), CpG islands, and gene annotations (Haeussler et al., 2019) are shown below to illustrate overlap between DMRs and functional elements in the genome.

B,C. Overlay of mCG signal for DMR regions shown in **A**.

D. Violin plot of mCG/CG level (top) and percent reduction, (DNMT3A$^{KO/+}$-WT)/WT, (bottom) in WT and DNMT3A$^{KO/+}$ cerebral cortex across indicated classes of genomic regions. Mean mCG/CG levels per-replicate are overlaid as dots on the violin plot. (n=4 per genotype; paired Student's T-Test with Bonferroni correction).

E. Heat map of CG DMRs called in the DNMT3A$^{KO/+}$ cortex. Biological replicates (B1-B4) are indicated.

F. Observed and background (from resampled DMRs, see *methods*) overlap between DNMT3A$^{KO/+}$ cortex CG-DMRs and various genomic regions (Fisher's Exact Test).

G. Violin plot of mCA/CA level (top) and percent reduction, (DNMT3A$^{KO/+}$-WT)/WT, (bottom) in WT and DNMT3A$^{KO/+}$ cerebral cortex across indicated classes of genomic regions. Mean mCA/CA levels per-replicate are overlaid as dots on the violin plot. (n=4; paired Student's T-Test with Bonferroni correction).

*, $P<0.05$; **, $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$. Bar graphs indicate mean and SEM across biological replicates.

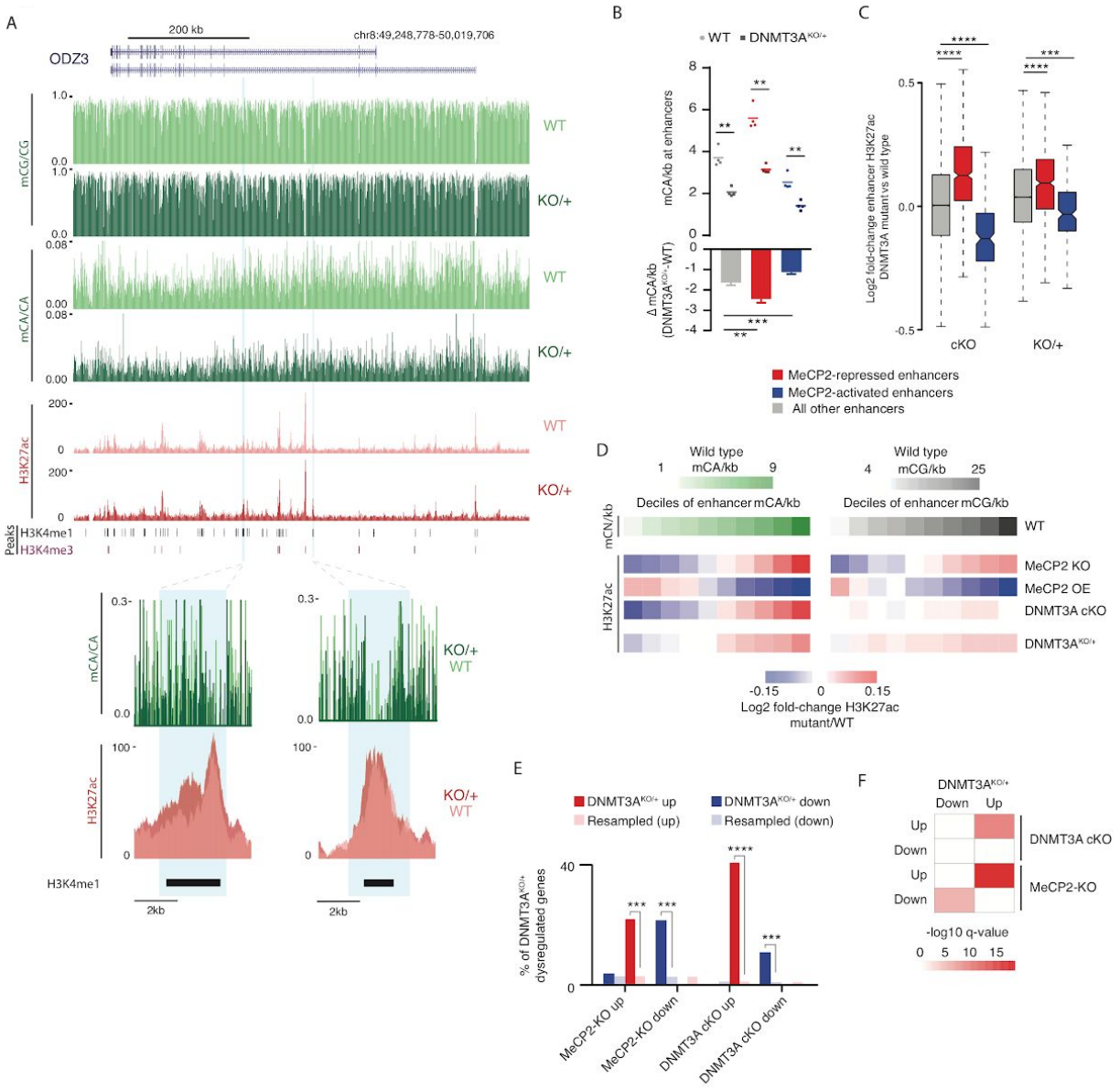

Figure 3.6. DNMT3A$^{KO/+}$ enhancer dysregulation and transcriptomic pathology overlaps with MeCP2 mutants.

A. Genome browser view of DNA methylation and H3K27ac ChIP-seq data from WT and DNMT3A$^{KO/+}$ cerebral cortex (top). Overlaid H3K27ac signal and mCA/CA levels at enhancer regions highlighted in blue that were identified as dysregulated enhancers upon disruption of mCA or MeCP2 (Clemens et al., 2019) (bottom).

B. Mean mCA sites/kb in WT and DNMT3A$^{KO/+}$ cortex (top) and number of mCA sites/kb lost in the DNMT3A$^{KO/+}$ cortex (bottom) for enhancers significantly dysregulated in

MeCP2 mutants (**, $P<0.01$; ***, $P<0.001$; n=4; paired Student's T-Test with Bonferroni correction).

C. Boxplot of fold-change in H3K27ac signal in DNMT3A *Baf53b*-cKO and the DNMT3A$^{KO/+}$ cortex for enhancers defined as significantly dysregulated in MeCP2 mutants (***, $P<10^{-8}$; **** $P<10^{-12}$; n=5 biological replicates of DNMT3A$^{KO/+}$ and WT; Wilcoxon test).

D. Heatmap of changes in H3K27ac signal for indicated mutants across deciles of enhancers sorted by wild-type mCA or mCG sites/kb ($P<2.2e^{-16}$, DNMT3A$^{KO/+}$ mCA/kb; $P<2.2e^{-16}$, DNMT3A$^{KO/+}$ mCG/kb; Spearman Rho correlation).

E. Observed versus background (estimated by resampling, see *methods*) overlap of significantly dysregulated genes (padj. < 0.1) in the DNMT3A$^{KO/+}$ and genes dysregulated in DNMT3A *Baf53b*-cKO or MeCP2 mutant mice (***, $P<e^{-5}$; ****, $P<e^{-10}$; hypergeometric test).

F. Significance of gene set expression changes in the indicated direction in the DNMT3A$^{KO/+}$ cortex for GAGE analysis of gene sets identified as dysregulated in DNMT3A *Baf53b*-cKO or MeCP2 mutant mice (Clemens et al., 2019).

Note: legend is shared in **B** and **C**. Box plots indicate median and quartiles. Bar graphs indicate mean and SEM.

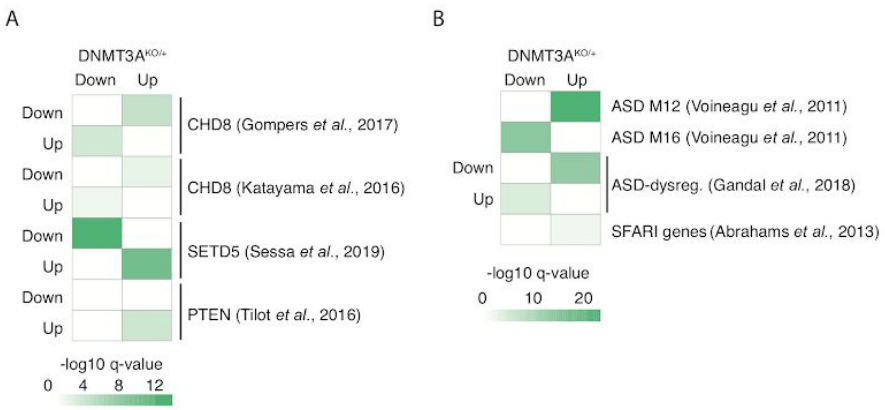

Figure 3.7. Gene dysregulation in the DNMT3A$^{KO/+}$ overlaps with other ASD/NDD disorders.

A. GAGE analysis of expression changes in DNMT3A$^{KO/+}$ for dysregulated gene sets identified in studies of NDD mouse models (Gompers et al., 2017; Katayama et al., 2016; Sessa et al., 2019; Tilot et al., 2016) (n=7 biological replicates of DNMT3A$^{KO/+}$ and WT).

B. GAGE analysis of expression changes in DNMT3A$^{KO/+}$ for gene sets identified in studies of human ASD. ASD module 12 (synaptic) and 16 (immune) were previously identified in weighted-gene coexpression analysis of human ASD brain (Voineagu et al., 2011), and ASD-dysregulated genes were previously identified (Abrahams et al., 2013; Gandal et al., 2018).

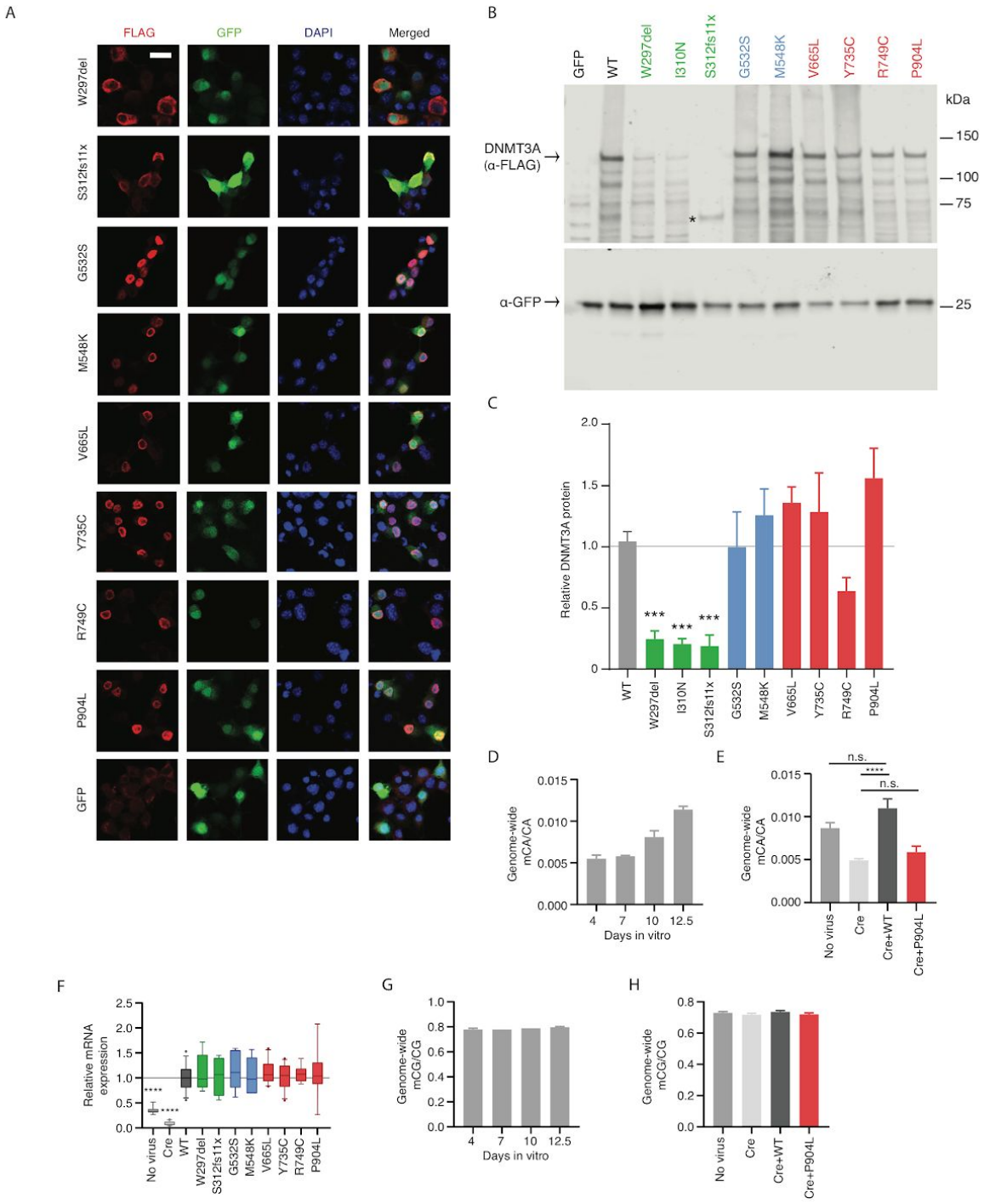

Figure 3.8 Related to Figures 3.1 and 3.2.

A. Example images of DNMT3A mutant protein immunocytochemistry in Neuro-2a cells. Scale bar = 20μm.

B. Full example immunoblot from Figure 1B with the truncated S312fs11x DNMT3A mutant protein indicated with an asterisk.
C. Quantification of immunoblot signal of DNMT3A (α-FLAG) from mutant proteins (n=6-30; unpaired Student's T-Test with Bonferroni correction). Though PWWP mutants show reduced protein expression, the remaining protein expression was adequate to assess cellular localization.
D. Genome wide mCA levels over time in WT neuronal cortical cultures, as measured by sparse WGBS (P=0.0035 effect by time, $F_{(3,4)}$=29.45, n=2; one-way ANOVA).
E. Genome-wide mCA levels in DNMT3A mutant add-back cortical cultures at DIV 12.5 (n=7-11; unpaired Student's T-Tests with Bonferroni correction).
F. qRT-PCR of *Dnmt3a* normalized to *Actb* for samples analyzed by WGBS. (n=4-11; one-sample Student's T-Test with Bonferroni correction).
G. Genome wide mCG levels over time in neuronal cortical cultures, as measured by sparse WGBS (P>0.05 effect by time, $F_{(3,4)}$=2.722, n=2; one-way ANOVA).
H. Genome-wide mCG levels in DNMT3A mutant add-back cortical cultures at DIV 12.5 (P>0.05; n=7-11; planned unpaired Student's T-Tests with Bonferroni correction).

***, P<0.001; ****, P<0.0001. Bar graphs indicate mean and SEM. Box plots contain 10^{th}-90^{th} percentiles of data, with remaining data represented as individual points.

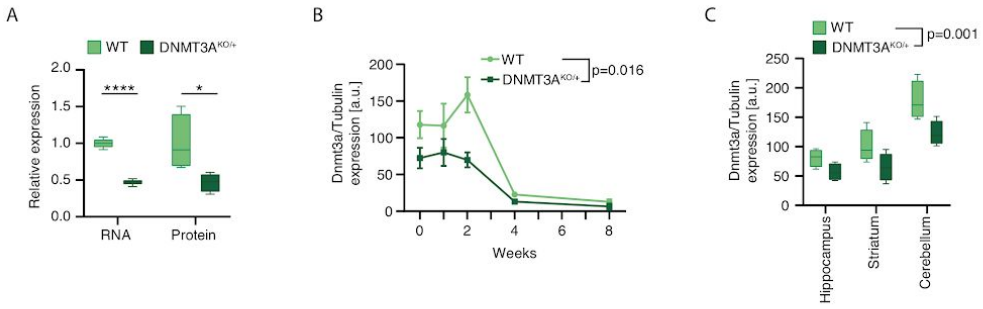

Figure 3.9 Related to Figures 3.3,3.4,3.5,3.6 and 3.7.

A. Normalized *Dnmt3a* mRNA and protein expression from 2-week cortices of DNMT3A$^{KO/+}$ and wild-type littermates (mRNA n=5, protein n=4, unpaired Student's T-Test).
B. Protein expression of DNMT3A normalized to α-Tubulin measured by western blotting for cerebral cortex of DNMT3A$^{KO/+}$ and wild-type littermates over developmental time (P=0.0157 effect by genotype, $F_{(1,16)}$=7.303, n=2-5; two-way ANOVA).
C. Protein expression of DNMT3A normalized to α-Tubulin measured by western blotting for hippocampus, striatum, and cerebellum (P=0.0010 effect by genotype, $F_{(1,18)}$=15.48, n=4; two-way ANOVA).

*, P<0.05; ****, P<0.0001. Line plots indicate mean and SEM. Box plots contain 10th-90th percentiles of data, with remaining data represented as individual points.

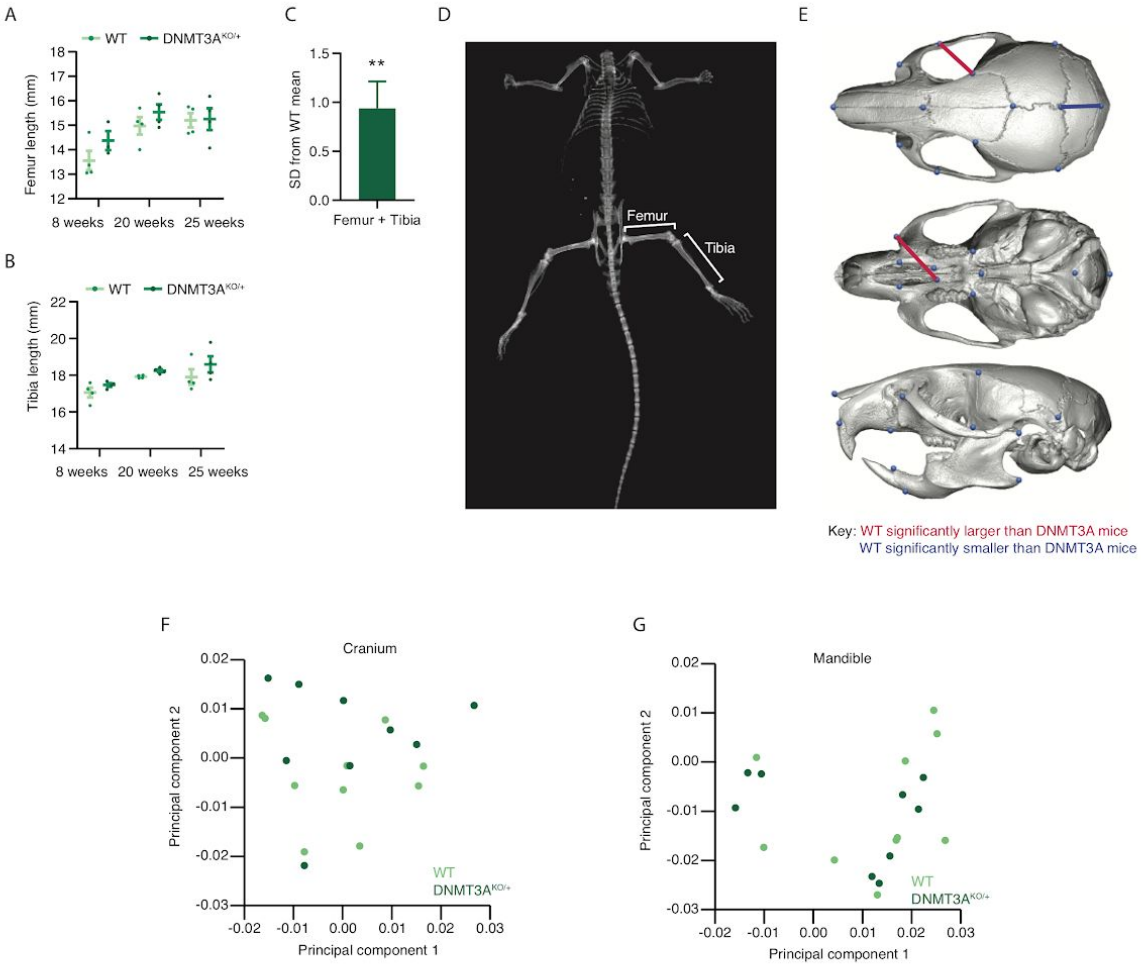

Figure 3.10 Related to Figure 3.3.

A,B Measurements of **A** femur and **B** tibia by dual x-ray imaging in WT and DNMT3A$^{KO/+}$ mice at 8, 20, and 25 weeks of age. Lines indicate mean and SEM, with points indicating individual animals.

C Standard deviations from WT mean for DNMT3A$^{KO/+}$ sum of femur and tibia. (n=12; paired Student's T-Test). Bar plot indicates mean and SEM.

D Example dual x-ray image of mouse body with femur and tibia indicated.

E Example of reconstructed skull from µCT imaging with landmarks used for craniofacial analysis shown. Red line indicates distances that are significantly larger in the WT compared to the DNMT3A$^{KO/+}$, while blue line indicates distance that is significantly smaller in the WT compared to the DNMT3A$^{KO/+}$ (P<0.05).

F,**G** Principal component analysis of **F** cranial and **G** mandibular shape shows no clear separation between groups along PC1 or PC2.

**, P<0.01. Bar graph indicates mean and SEM.

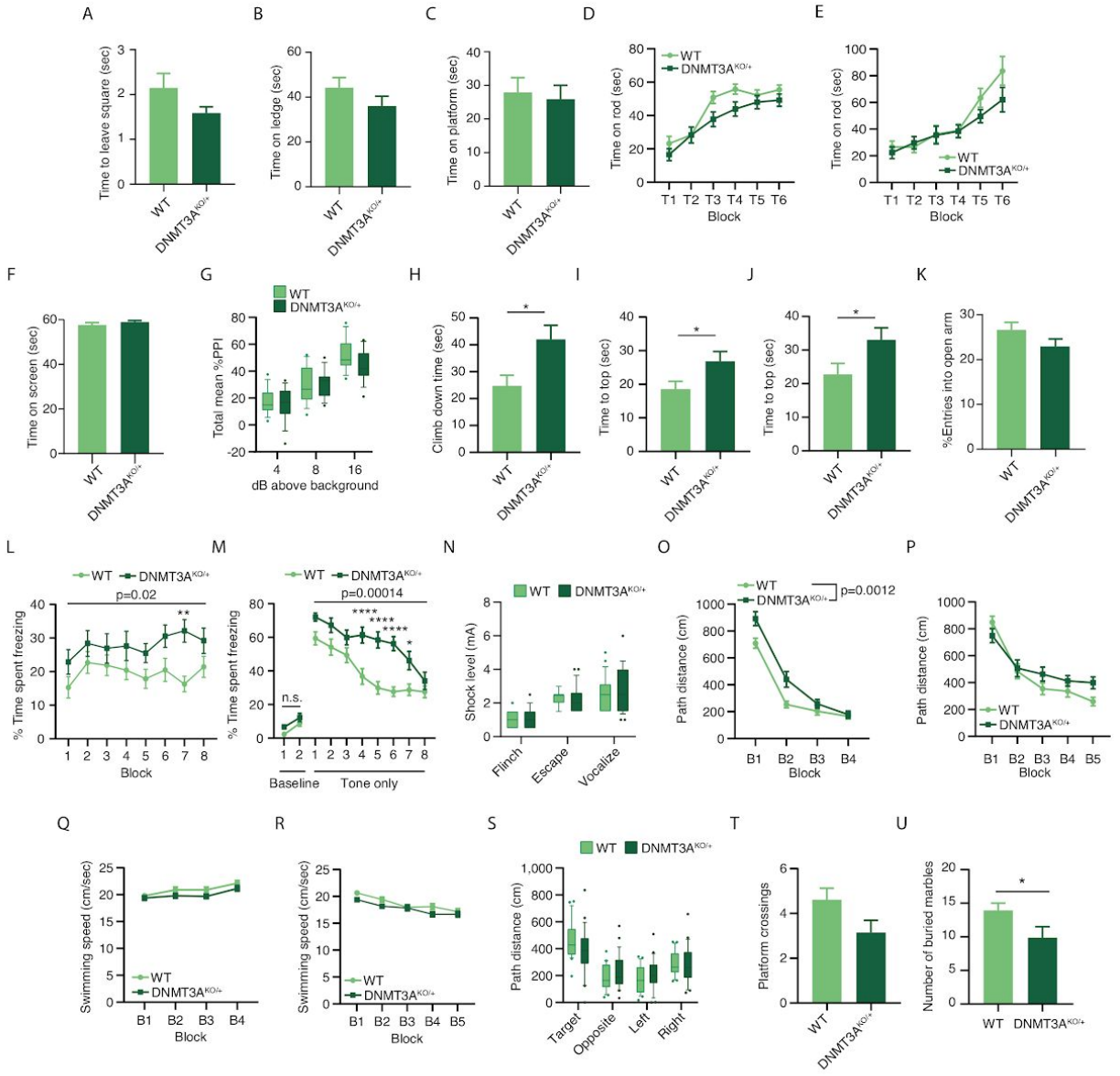

Figure 3.11 Related to Figure 3.3.

(**A-J**) Comparison of DNMT3A$^{KO/+}$ and WT mice across a battery of sensorimotor assays. DNMT3A$^{KO/+}$ mice show no significant difference in (**A**) walking initiation, or latency to fall off (**B**) ledge or (**C**) platform. (**D-E**) DNMT3A$^{KO/+}$ mice show no difference compared to WT littermates in motor coordination as evidenced by time on a continuous (**D**) and accelerating (**E**) rotarod. (**F**) DNMT3A$^{KO/+}$ mice show no difference in grip strength compared to WT littermates evidenced by no change in time on an inverted screen. (**G**) Mean % pre-pulse inhibition shows no significant difference between genotypes. DNMT3A$^{KO/+}$ mice show a significant increase in time to (**H**) climb down a pole (P=0.016, n=21,27; unpaired Student's T-Test), and to the top of a (**I**) 60° inclined screen (P=0.039, n=21,27; unpaired Student's T-Test) and a (**J**) 90° inclined screen (P=0.045, n=21,27; unpaired Student's T-Test). (**K**) DNMT3A$^{KO/+}$ mice show no deficit in elevated plus maze exploration as measured by percent entries into open arms. (**L**) Percent

time spent freezing in contextual fear trials (P=0.0215 effect by genotype, $F_{(1,50)}$=5.633, n=26; two-way repeated-measures ANOVA), and (**M**) cued fear trials (Baseline: P=0.0606 effect by genotype, $F_{(1,50)}$=3.685; Cue: P<0.0001 effect by genotype, $F_{(1,50)}$=17.03; n=26; two-way repeated-measures ANOVA with Sidak's multiple comparisons test). (**N**) Shock sensitivity during conditioned fear test as indicated by the minimum shock needed to exhibit a behavioral response in mice shows no significant difference between genotypes. (**O-R**) Path distance to escape platform and swim speeds in the Morris water maze task. DNMT3A$^{KO/+}$ mice show increased path distance to escape platform in both (**O**) cued trials (P=0.0012 effect by genotype, $F_{(1,46)}$=11.93; P=0.0433 interaction effect of genotype and trial block, $F_{(3,138)}$=2.784; n=21,27; two-way repeated-measures ANOVA) and (**P**) place trials (P=0.0408 interaction effect of genotype and trial block, $F_{(4,184)}$=2.55 n=21,27; two-way repeated-measures ANOVA). No significant difference is seen in swimming speed during (**Q**) cued trials (P=0.0634 effect by genotype, $F_{(1,46)}$=3.619 n=21,27; two-way repeated-measures ANOVA) and (**R**) place trials (P=0.098 effect by genotype, $F_{(1,46)}$=2.845, n=21,27; two-way repeated-measures ANOVA). (**S**) DNMT3A$^{KO/+}$ mice show no significant difference in time spent in the target quadrant of a Morris water maze compared to WT littermates. (**T**) DNMT3A$^{KO/+}$ mice show a trend towards a reduction in platform crossings in the probe trial (P=0.0609; n=21,27; unpaired Student's T-Test). (**U**) After 30 minutes, DNMT3A$^{KO/+}$ mice have significantly fewer total marbles buried (P=0.0438; n=13,14; unpaired Student's T-Test).
*, P<0.05; **, P<0.01; ****, P<0.0001. Bar graphs and line plots indicate mean and SEM. Box plots contain 10th-90th percentiles of data, with remaining data represented as individual points.

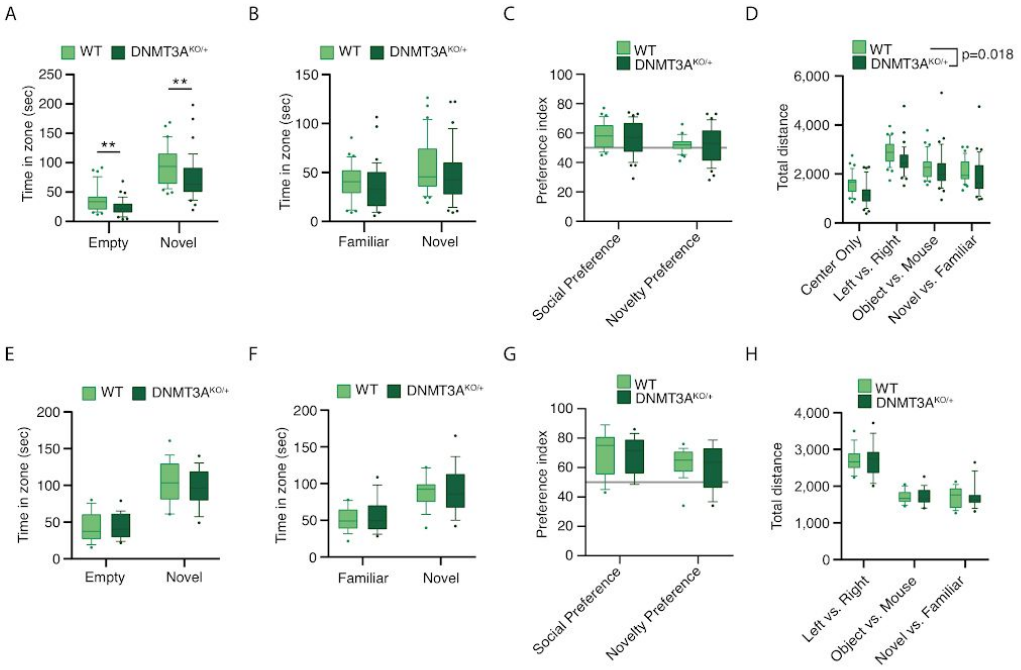

Figure 3.12 Related to Figure 3.3.

A-D Social approach results from moderate-light apparatus show reduced activity in DNMT3A$^{KO/+}$ mice.

A Quantification of time spent in zones closest to each cup during phase with empty cup and novel conspecific (Empty, *P*=0.0026; Novel, *P*=0.0095; n=33,39; unpaired Student's T-Test).

B Time spent in zones closest to novel mouse and familiar mouse (Familiar, *P*=0.29; Novel, *P*=0.24; n=33,39; unpaired Student's T-Test).

C Preference Index for novel conspecific in the 3-chambered social approach task for WT and DNMT3A$^{KO/+}$ mice using time spent in zones as calculated: Mouse/(Mouse+Object)x100 or Novel/(Novel+Familiar)x100 within each animal.

D Total distance traveled during the 3-chambered social approach task for WT and DNMT3A$^{KO/+}$ mice shows a broad reduction of distance traveled across all trials by DNMT3A$^{KO/+}$ mice (*P*=0.018 effect by genotype, $F_{(1,70)}$=5.862, n=33,39; two-way ANOVA).

E-H Social approach results from low-light apparatus with minimal experimenter presence show no differences between wild-type and DNMT3A$^{KO/+}$ mice.

E Quantification of time spent in zones closest to each cup during phase with empty cup and novel conspecific (Empty, *P*=0.85 Novel, *P*=0.49; n=19,18; unpaired Student's T-Test).

F Time spent in zones closest to novel mouse and familiar mouse (Familiar, *P*=0.75; Novel, *P*=0.86; n=19,18; unpaired Student's T-Test).

G Preference Index for novel conspecific in the 3-chambered social approach task shows no change for WT and DNMT3A$^{KO/+}$ mice calculated using the same approach as C.

H Total distance traveled during the 3-chambered social approach task for WT and DNMT3A$^{KO/+}$ mice shows no reduction of distance traveled across trials by DNMT3A$^{KO/+}$ mice (P=0.77 effect by genotype, $F_{(1,105)}$=0.08, n=19,18; two-way ANOVA).

**, P<0.01. Box plots contain 10th-90th percentiles of data, with remaining data represented as individual points.

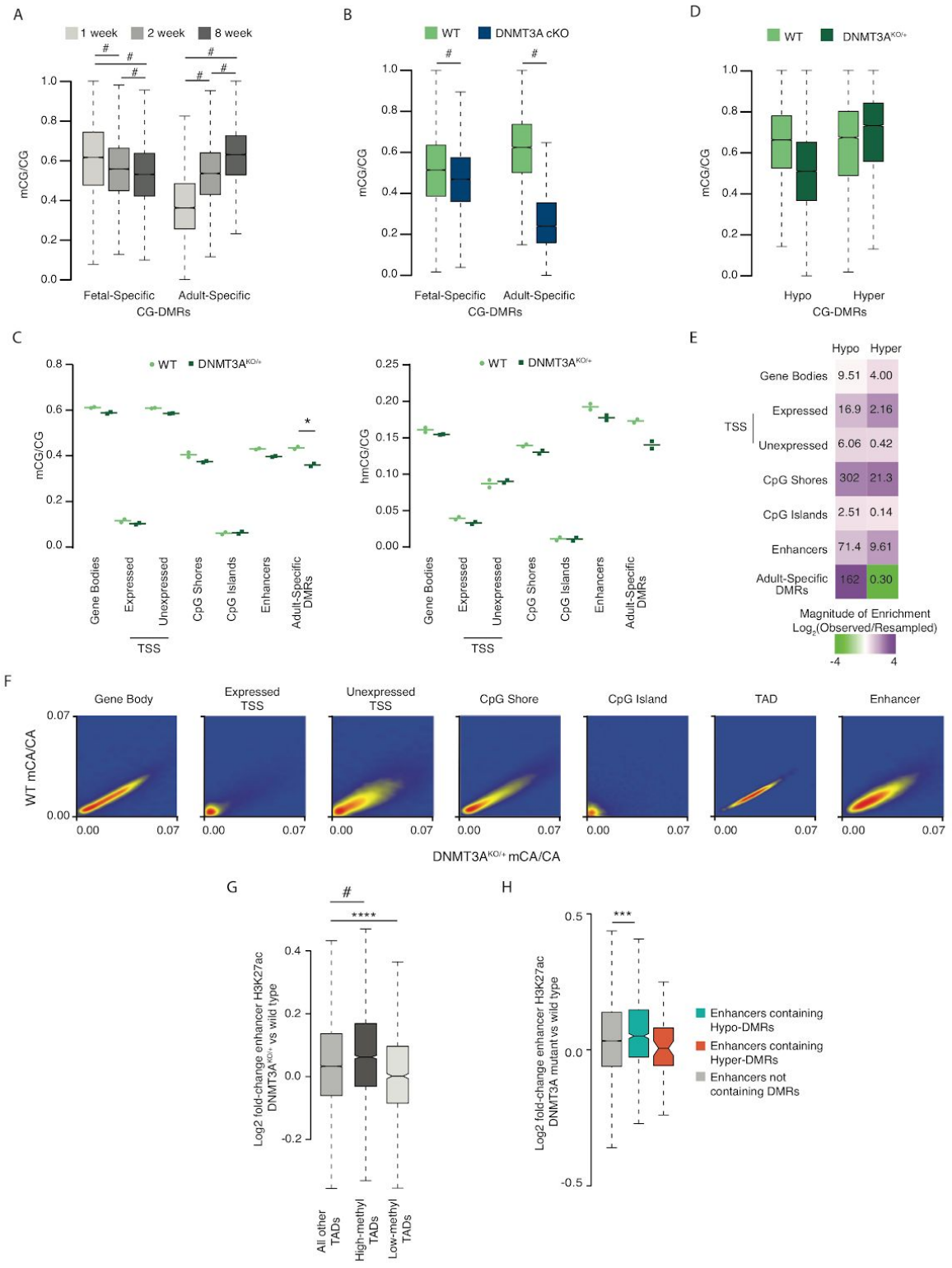

Figure 3.13 Related to Figures 3.4,3.5, and 3.6.

A. Boxplots of CG methylation over postnatal development for regions called as having higher methylation in the frontal cortex of fetal versus adult tissue (Fetal-Specific DMR),

or called as having higher methylation in the frontal cortex of the adult versus fetal tissue (Adult-specific DMR). DMRs from Lister et al. 2013, methylation data from Stroud et al. 2017. (Wilcoxon rank sum test with Bonferroni correction)

B. Boxplots of cortical methylation of WT and DNMT3A *Nestin*-cKO (Stroud et al. 2017) at 8 weeks postnatal within developmental DMRs shown in **A** (Wilcoxon rank sum test with Bonferroni correction).

C. CG/CG (left) and hmCG/CG (right) from WT and DNMT3A$^{KO/+}$ cortices in various genomic contexts as measured by oxidative bisulfite sequencing (paired Student's T-Test with Bonferroni correction).

D. Boxplots of cortical methylation of WT and DNMT3A$^{KO/+}$ at 8 weeks postnatally within DMRs defined in the DNMT3A$^{KO/+}$ model. DMRs are called on this data set, so no additional statistics were run on genotype differences.

E. Heatmap of log2 odds-ratios of observed and background (derived from resampling, see *methods*) overlap between DNMT3A$^{KO/+}$ cortex CG-DMRs and various genomic regions. Color indicates magnitude of enrichment (Observed/resampled), number indicates -log10 p-value for Fisher-exact tests between observed and resampled values.

F. Smooth scatter plots of WT and DNMT3A$^{KO/+}$ mCA/CA for classes of genomic regions.

G. Boxplot of fold-changes in H3K27ac signal in the DNMT3A$^{KO/+}$ and WT cortex for enhancers found within TADs with high (top 10%) and low (bottom 10%) average levels of mCA/CA.

H. Boxplot of fold-changes in H3K27ac signal in the DNMT3A$^{KO/+}$ and WT cortices from enhancers containing hypo/hyper CG-DMRs, and from enhancers containing no DMRs.

***, $P<0.001$; ****, $P<0.0001$; #, $P<2.2e^{-16}$. Box plots indicate median and quartiles.

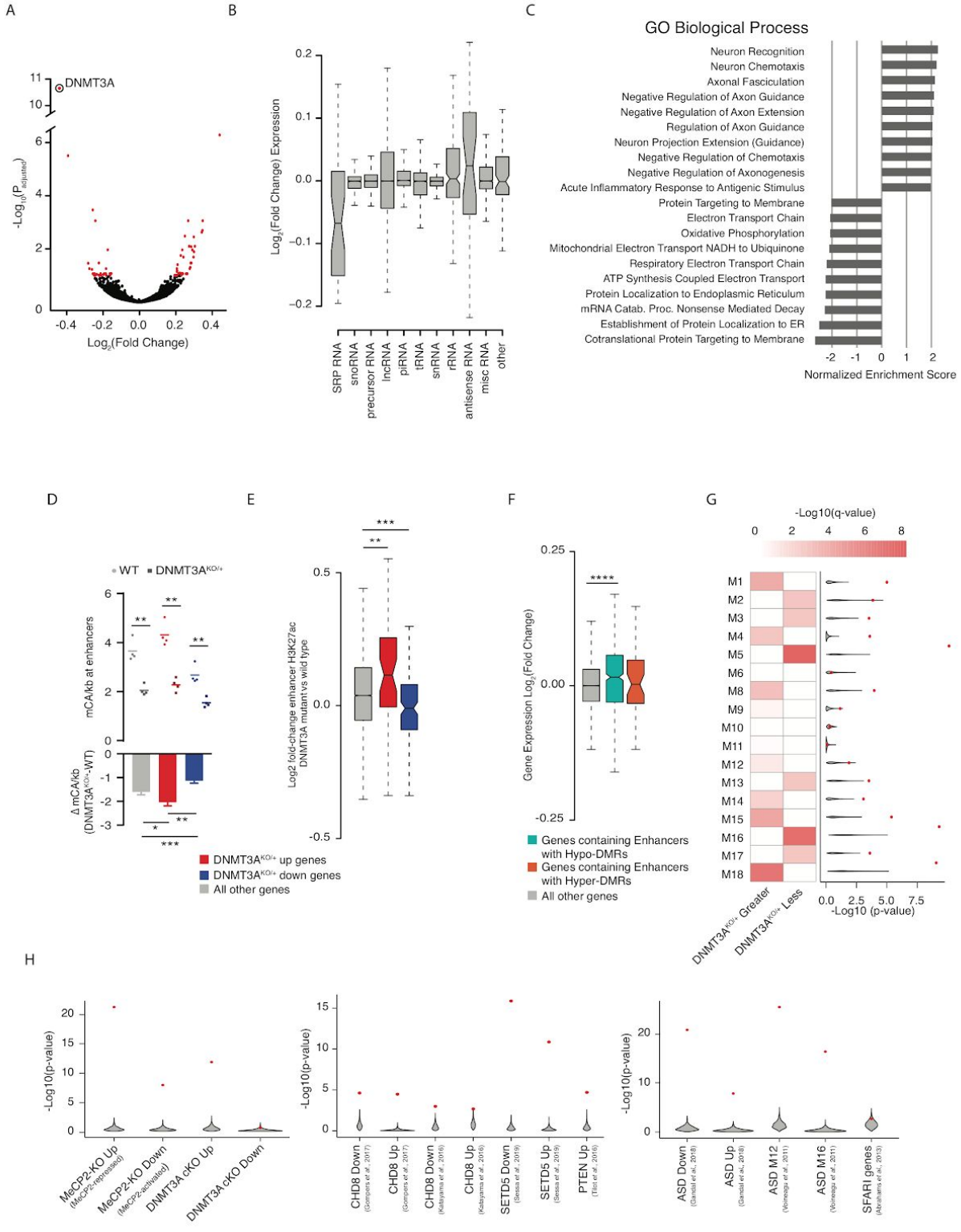

Figure 3.14 Related to Figures 6 and 7.

A. Volcano plot of DESeq $\log_2$ fold changes of the DNMT3A$^{KO/+}$ versus WT. Genes reaching a significance of $p_{adj.}$<0.1 are colored in red.
B. Boxplot of DESeq $\log_2$ fold changes from noncoding RNAs from RNAcentral, organized by RNA type.
C. Top ten up- and down-regulated Gene Ontology terms from Broad GSEA Molecular Signatures Database version 7.0 (Subramanian et al., 2005). All terms are significant at an FDR<0.1.
D. Mean mCA sites/kb in WT and DNMT3A$^{KO/+}$ cortex (top) and number of mCA sites/kb lost in the DNMT3A$^{KO/+}$ cortex (bottom) for enhancers found within genes called as significantly dysregulated in the DNMT3A$^{KO/+}$ (n=4; paired Student's T-Test with Bonferroni correction).
E. Boxplot of fold-change in H3K27ac signal in enhancers found within genes called as significantly dysregulated in the DNMT3A$^{KO/+}$ (n=5 biological replicates of DNMT3A$^{KO/+}$ and WT; Wilcoxon test)
F. Boxplot of DESeq $\log_2$ fold changes of in gene expression in the DNMT3A$^{KO/+}$ versus WT, for genes containing enhancers with hypo-CG-DMRs, genes containing enhancers with hyper-CG-DMRs, and all other genes containing enhancers.
G. GAGE analysis of developmental expression modules (Parikshak et al., 2013). Significant modules (q-value<0.1) are colored in red (left). Expression matched resampling of each gene set was performed 1,000 times and analyzed using GAGE for enrichment in DNMT3A$^{KO/+}$ fold-change data (gray violin). This was compared with the true gene set p-value (red point) to test for significance (right). Only the direction of dysregulation in which the gene sets showed the highest significance (i.e. DNMT3A$^{KO/+}$ greater or less) is shown.
H. Expression matched resampling of GAGE analysis for gene sets displayed in Figures 6 and 7. Only the direction of dysregulation in which the gene set showed significance (i.e. DNMT3A$^{KO/+}$ greater or less) is shown.

Note: legend is shared in **D** and **E.** *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001.

Chapter 4: Conclusions and future directions

This dissertation begins to interrogate the influences on mCA deposition genome-wide, and how the megabase scale of topologically associating domains enables MeCP2 to repress many enhancers within highly methylated domains. As these enhancers preferentially contact the genes they are in, repression of these enhancers represses long, highly methylated genes that are likely to contain multiple enhancers (Clemens et al. 2020). This method of neural regulation may apply to multiple disease models, as we find that a mouse heterozygous knockout of DNMT3A (modeling TBRS) exhibits a 50% decrease in mCA with minor effects on mCG (Christian et al. 2020). These methylation reductions contribute to overlapping gene expression and enhancer activity changes with the MeCP2 knockout, as MeCP2 is likely less able to bind to what were previously highly methylated enhancers.

In preliminary research I have leveraged a library of existing ChIP-seq and bisulfite-seq data (Stroud et al., 2017) in order to interrogate the exact factors that affect mCA deposition in the genome. My work first verified that the observed pattern of a negative correlation between gene expression and that gene's mCA levels. However, this relationship is known to be weak, with many highly expressed yet highly methylated genes, and lowly expressed, lowly methylated genes throughout the genome. Correlations between genic mCA and expression are also relatively low, with a Pearson r value of -0.18. Previous work noted that the methylation of the surrounding TAD can have large impacts on gene methylation (Clemens et al., 2020), so I used

both factors to build a linear model on both expression and TAD methylation (Figure 4.1). This model showed significant improvement over using either TAD methylation or expression alone (Pearson $r^2 = 0.33809$). In addition, the model demonstrated a lowly significant interaction term between TAD methylation and gene expression, indicating that the expression-associated demethylation effect occurs independently from the amount of TAD methylation around the gene.

While the expression-mCA inverse relationship has been noted in the literature (Stroud et al. 2017), less is known regarding the specific factors that might be associated with gene expression or TAD methylation that might impact genic mCA. For example, H3K36me3 within gene bodies is associated with expression (Kolasinska-Zwierz et al., 2009), but the presence of phosphoserine-5 polymerase II is the most direct indication of active transcription through the region (Komarnitsky et al. 2000; Ahn et al. 2004; Phatnani and Greenleaf 2006). Thus, it is important to determine exactly which ChIP-seq signals and histone marks may have a direct effect on DNMT3A binding and mCA deposition within the gene body. To this end, I analyzed a variety of ChIP-seq signals generated in 2-week postnatal mouse cortex (Stroud et al., 2017) and tested the performance of linear models on all pairwise comparisons of each of the ChIP marks (Figure 4.2). This analysis found that the predictive power of each variable was relatively dispersed, but several factor-pairs consistently resulted in better models than other signals. Phosphoserine Pol II generated well-performing models with a variety of paired signals, for instance, but H3K36me3 did not add much accuracy to many variables, suggesting that the relationship between expression and mCA is not mediated by H3K36me3 presence.

Interestingly, H3K9me2, while not generally well-predicting, makes the most predictive model of the set specifically when combined with H3K27me2, suggesting that the H3K9me2 relationship to mCA is dependent on the amounts of H3K27me2 within the region.

The analysis of these ChIP signals indicated that DNMT3A binding could not be predicted solely by a few biological signals, and that the relationship between these factors and mCA deposition may be more complicated than can be modeled linearly. To circumvent these limitations, I trained a random forest (RF) model with 5-fold cross-validation (Liaw et al. 2002; Stone 1974; Dietterich 1998) on all ChIP-seq marks previously used in my model. This approach successfully predicted mCA patterns with high accuracy (Pearson $r^2 = 0.685911$), and feature importances roughly matched with the performance of each factor in linear modeling, with the most important factors being phosphoserine Pol II, and H3K27me2. Because it had been previously noted that the methylation of extragenic regions depended on different factors than the methylation of intragenic regions (Stroud et al., 2017), I also trained and tested a random forest model on extragenic data, as well as both extragenic and intragenic data. Many features between intragenic and extragenic data shared similar amounts of importance, but the most important factors between extragenic and intragenic regions were generally not shared. H3K27me2 and phosphoserine Pol II had minimal importance in the extragenic data, which had H3.3 as the best predicting mark directing methylation. This result may be driven by a bimodal relationship of mCA and H3.3, as there is a high correlation of H3.3 to mCA within the low levels of H3.3 in extragenic regions, but a negative correlation when H3.3 reaches higher levels within gene bodies (figure 4.3).

While the predictive power of the random forest model is higher than what could be achieved from a linear model or from any single factor individually, there still remains a large amount of variance unexplained (~30%). This gap could be due to noise in the measurement of mCA or ChIP signal, but it could also be due to a lack of relevant biological information, such as a chromatin mark or protein that recruits DNMT3A that wasn't assessed in the original set of factors used. A particularly relevant factor to assess and test is H3K36me2, which has recently been suggested to bind DNMT3A in non-neural cells (Weinberg et al., 2019). In addition to the recent work linking DNMT3A and H3K36me2, mutations in NSD1, a histone methyltransferase responsible for depositing H3K36me2, cause Sotos syndrome, a disease with overlapping features with TBRS (Kurotaki et al., 2002). Further linking Sotos syndrome to mCA is evidence that methylation in general is disrupted in Sotos syndrome patient blood (Choufani et al., 2015). To examine if H3K36me2 occupancy could be used to predict mCA levels, we generated H3K36me2 ChIP-seq data from 2-week cortex, and I used this data to predict mCA, along with the other data generated previously. The results were somewhat surprising, as although H3K36me2 was the most important variable reported in a random forest model, its addition did not give the model any noticeable improvement in accuracy. This may be an effect of the high amount of noise in this additional assay, or an indication that the information that would be added by H3K36me2 could be interpolated from the many other factors used in the model.

In order to define a minimal set of factors that could accurately predict mCA, and to see if H3K36me2 would be included in this minimal set, I implemented a recursive feature addition

approach that assessed the performances of models given different numbers of features. This approach found that 6-8 factors were required to predict mCA with 96-99% of the power of the full model, highlighting their importance to DNMT3A and mCA deposition. Further research focused on these factors has the potential to uncover critical regulators of mCA patterning across the genome.

Given the known links between H3K36me2 and DNMT3A, the overlapping phenotypes exhibited in disruptions of H3K36me2 and mCA establishers, and the degree to which mCA patterns rely on this histone mark, there is a pressing need to verify the link between H3K36me2 and mCA. Research into the relationship between NSD1 and the neuronal epigenome could also shed light into the mechanisms of Sotos syndrome and whether it is affected by alterations in mCA. To assess the effects of H3K36me2 loss on mCA, we are planning to generate a knockout of NSD1 within neurons and assess the transcriptomic and epigenomic effects this has within the brain. We predict that, unlike mutations of DNMT3A, that global mCA levels in the brain will be largely unchanged. However, it is likely that the patterns and distribution of mCA throughout the genome will be altered, and regions rich in H3K36me2 in wildtype conditions will have substantially lower methylation levels in the NSD1 knockout. Enhancers and genes within these newly demethylated regions would then be expected to increase in activity and expression, overlapping with effects observed in the DNMT3A and MeCP2 mutants. If we observe these effects, it will establish a previously unrecognized function for H3K36me2 in guiding mCA deposition, and provide direct evidence that mCA may be disrupted in another neurodevelopmental disorder, Sotos syndrome.

To further test the biological significance of the minimal set of mCA-predicting factors identified by machine learning approaches, we are also planning on assessing these marks in a separate region of the brain, the cerebellum, and applying a random forest model, previously trained on these marks within the cortex. The accuracy of these predictions, versus the accuracy of predictions made on a model trained on cerebellar data, will then tell us if the relationships between mCA and these different biological variables are consistent between brain regions. This approach would enable us to computationally verify the biological importance of these factors without requiring a lengthy biochemical assay or generating numerous knockout animals.

In conclusion, my dissertation work finds that a small number of factors are predictive of neuronal mCA deposition, and are likely influenced by TADs, which are known to be regions of consistency of many histone modifications. This TAD-level of variance then establishes ~100-kilobase sized regions of enriched mCA, which have the potential to contain tens or hundreds of enhancers and other regulatory elements. MeCP2 can bind and repress these enhancers en-masse in order to repress long, highly-methylated genes, which are preferentially linked to and activated by the many highly-methylated enhancers that they contain. This mechanism of repression by MeCP2 may apply to multiple neurodevelopmental disorders, each of which may alter different steps in the mCA establishment-readout pathway, generating different, but overlapping transcriptomic, epigenomic, and phenotypic features. More research into the proteins that surround and interact with this pathway has the potential to shed insight into a variety of neurodevelopmental disorders, as they may be linked to mCA-based regulation.

4.1 Figures

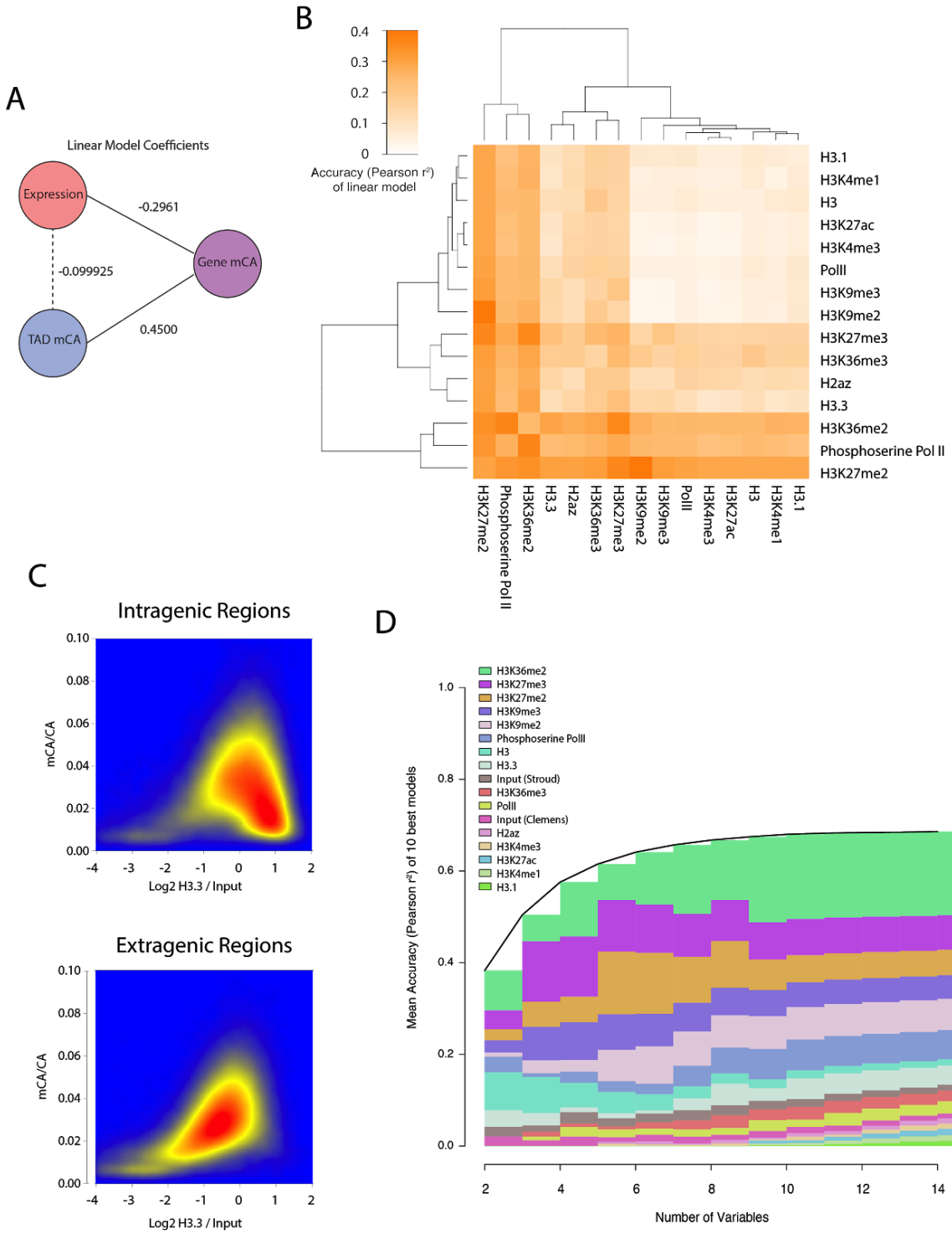
A
Linear Model Coefficients
Expression
Gene mCA
TAD mCA
-0.2961
-0.099925
0.4500
B
Accuracy (Pearson r²) of linear model
H3.1
H3K4me1
H3
H3K27ac
H3K4me3
PolII
H3K9me3
H3K9me2
H3K27me3
H3K36me3
H2az
H3.3
H3K36me2
Phosphoserine Pol II
H3K27me2
C
Intragenic Regions
Extragenic Regions
mCA/CA
Log2 H3.3 / Input
D
Mean Accuracy (Pearson r²) of 10 best models
Number of Variables
Input (Stroud)
Input (Clemens)

Figure 4.1. Machine learning approaches find a minimal set of mCA predictive factors

A. Conceptual illustration of linear model combining expression and TAD mCA levels (mC/CA) to predict genic mCA levels. Lines linking components of the linear model are labeled with their weighted coefficients. Accuracy of final model fit is Pearson r^2 = 0.33809
B. Heatmap of r^2 values for linear models using ChIP-seq signals to fit mCA levels at genic regions, given combinations of 2 factors. Diagonal of matrix is the r^2 value of a linear model given only that single factor. Models built with 5kb intragenic bins of the genome, placed at least 3kb away from transcriptional start sites.
C. Density scatter plots of input-normalized H3.3 ChIP-seq signal and mCA levels at 5kb regions across the genome, either within or outside of genes.
D. Line plot, plus bar plots, of the mean accuracy and feature importances of random forest prediction models, for increasing numbers of variables. Only the 10 best performing combinations of variables are plotted. Mean feature importances are calculated from their feature importances in each of the 10 best performing combinations of variables - a model that does not use that feature is considered to be an importance of 0. Mean feature importances are scaled to the performance of the models used, for readability, and all have total sums of importances of 1, prior to scaling.

All analysis shown was carried out for mouse cerebral cortex. ChIP-seq data were generated from 2-week postnatal mice, and bisulfite-seq data is from 8-week postnatal cortex. H3K36me2 ChIP-seq data was generated for this study, and the rest of the dataset was generated by (Stroud et al. 2017)

4.2 References Cited

Ahn, Seong Hoon, Minkyu Kim, and Stephen Buratowski. 2004. "Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3' End Processing." *Molecular Cell* 13 (1): 67–76.

Choufani, S., C. Cytrynbaum, B. H. Y. Chung, A. L. Turinsky, D. Grafodatskaya, Y. A. Chen, A. S. A. Cohen, et al. 2015. "NSD1 Mutations Generate a Genome-Wide DNA Methylation Signature." *Nature Communications* 6 (December): 10207.

Christian, Diana L., Dennis Y. Wu, Jenna R. Martin, J. Russell Moore, Yiran R. Liu, Adam W. Clemens, Sabin A. Nettles, et al. 2020. "DNMT3A Haploinsufficiency Results in Behavioral Deficits and Global Epigenomic Dysregulation Shared across Neurodevelopment Disorders." *bioRxiv*. https://doi.org/10.1101/2020.07.10.195859.

Clemens, Adam W., Dennis Y. Wu, J. Russell Moore, Diana L. Christian, Guoyan Zhao, and Harrison W. Gabel. 2020. "MeCP2 Represses Enhancers through Chromosome Topology-Associated DNA Methylation." *Molecular Cell* 77 (2): 279–93.e8.

Dietterich, T. G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation* 10 (7): 1895–1923.

Kolasinska-Zwierz, Paulina, Thomas Down, Isabel Latorre, Tao Liu, X. Shirley Liu, and Julie Ahringer. 2009. "Differential Chromatin Marking of Introns and Expressed Exons by H3K36me3." *Nature Genetics* 41 (3): 376–81.

Komarnitsky, Philip, Eun-Jung Cho, and Stephen Buratowski. 2000. "Different Phosphorylated Forms of RNA Polymerase II and Associated mRNA Processing Factors during Transcription." *Genes & Development* 14 (19): 2452–60.

Kurotaki, Naohiro, Kiyoshi Imaizumi, Naoki Harada, Mitsuo Masuno, Tatsuro Kondoh, Toshiro Nagai, Hirofumi Ohashi, et al. 2002. "Haploinsufficiency of NSD1 Causes Sotos Syndrome." *Nature Genetics* 30 (4): 365–66.

Liaw, Andy, Matthew Wiener, and Others. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Phatnani, Hemali P., and Arno L. Greenleaf. 2006. "Phosphorylation and Functions of the RNA Polymerase II CTD." *Genes & Development* 20 (21): 2922–36.

Stone, Mervyn. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 36 (2): 111–33.

Stroud, Hume, Susan C. Su, Sinisa Hrvatin, Alexander W. Greben, William Renthal, Lisa D. Boxer, M. Aurel Nagy, et al. 2017. "Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States." *Cell* 171 (5): 1151–64.e16.

Weinberg, Daniel N., Simon Papillon-Cavanagh, Haifen Chen, Yuan Yue, Xiao Chen, Kartik N. Rajagopalan, Cynthia Horth, et al. 2019. "The Histone Mark H3K36me2 Recruits DNMT3A and Shapes the Intergenic DNA Methylation Landscape." *Nature* 573 (7773): 281–86.