

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Functional characterization of neuronal cis-regulation

Permalink

<https://escholarship.org/uc/item/6vq871rz>

Author

Laboy Cintron, Dianne

Publication Date

2024

Supplemental Material

<https://escholarship.org/uc/item/6vq871rz#supplemental>

Peer reviewed|Thesis/dissertation

Functional characterization of neuronal cis-regulation

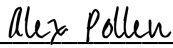
by
Dianne Laboy Cintron

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in
Biomedical Sciences


in the
GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

2102D531BF2D416... Alex Pollen
Chair

DocuSigned by:

Nadav Ahituv

DocuSigned by:

0C4570A6FDE94F2... DEVANAND MANOLI

Committee Members

Copyright 2024

by

Dianne Laboy Cintrón

ALL RIGHTS RESERVED

DEDICATION

I dedicate my dissertation work to my parents, Isabel Cintrón Carrasquillo and David Laboy Reyes, who made many sacrifices to provide me with the best education.

Esta tesis está dedicada a mis padres, Isabel Cintrón Carrasquillo y David Laboy Reyes, quienes sacrificaron mucho para darme la mejor educación.

ACKNOWLEDGEMENTS

First, I would like to thank my scientific advisor and mentor, Dr. Nadav Ahituv. I first learned about the research of Dr. Nadav Ahituv as an undergraduate. I was fascinated by Dr. Ahituv's approach to science through his amazing publication record in a variety of scientific disciplines. Rotating in his lab at the beginning of 2020 affirmed me that not only was the research exciting but he had cultivated a supportive laboratory environment. I joined the Ahituv Lab during March of 2020, at a time of uncertainty our weekly meetings were a time that allowed me to focus on my doctoral journey. Dr. Ahituv's optimism throughout failed experiments and mistakes, gave me the certainty that I could accomplish this degree. He supported me throughout the most challenging times during graduate school. Finally, Dr. Ahituv's mentorship and training allowed me to develop into the scientist I am today. I am forever thankful and will carry the lessons learned forward in my scientific journey.

I would also like to extend a thank you to other UCSF faculty. First, I want to thank my thesis committee, Dr. Alex Pollen and Dr. Devanand Manoli for their support, advise and expertise. I entered our meetings feeling like I never had done enough. However, I left our meetings encouraged on my scientific accomplishments and progress. Second, I would like to extend a thank you to other members of my qualifying exam committee, Dr. Licia Selleri and Dr. Allison Xu. Finally, I want to thank Dr. Jason Sello for all the advice and guidance he provided as IMSD mentoring faculty.

Next, I would like to extend my gratitude to the members of the Ahituv Lab. I would like to extend a thank you to the group that welcomed me into the lab: Dr. Serena Tamura, Dr. Lana Harshman, Dr. Wei Gordon and Rachael Bradley. I will forever look

up to this amazing group of women. During my graduate school journey, it felt like I had four older sisters who paved the way to make my journey better. I would like to acknowledge the mentorship and help from the amazing post-doctoral fellows and visiting scholars in the lab: Dr. Ofer Yizhar Barnea, Dr. Jamal Ghoumid, Dr. Yelena Guttman, Dr. Coline Arnold, Dr. Aki Ushiki, Dr. Sarah Fong Dr. Chengyu Deng, and Dr. Yarden Golan. I also want to extend a thank you to other lab members which helped me throughout my time in the lab: Dr. Xujia Zhou, Mai Nobuhara, Ryder Easterlin and Rory Sheng. Finally, I want to extend a thank you to Candace Chan for all the chats, honest feedback, and deadlines reminders. At times graduate school can be isolating but the amazing people from the Ahituv Lab made sure I was always supported. I am forever thankful for all the support, laughs, discussions, and moments that made my graduate research experience one to fondly look back at.

Next, I want to acknowledge my previous scientific mentors and advisors: Dr. Takato Imaizumi, Dr. Akane Kubota, Dr. Heather Mefford, Dr. Allison Muir and Dr. Scott Freeman. My first lab experience was in the Imaizumi lab where Dr. Akane Kubota taught me the fundamentals of how to conduct scientific research. I am indebted to Dr. Kubota for all her patience and trust as my first scientific mentor. After my time in the Imaizumi Lab, I joined the Mefford Lab to explore my interest in genetics. Dr. Heather Mefford and Dr. Allison Muir taught me so much about human genetics, clinical genetics, and cell culture. I am so thankful to have been part of the Mefford Lab where I gained the expertise to be able to do independent research in graduate school. Finally, I want to thank my biology professor, Dr. Scott Freeman. I am so thankful for all the work Dr. Freeman has done towards inclusive teaching in biology and for letting me take part

of his research. My undergraduate scientific mentors were key in my development as a researcher and their encouragement allowed me to feel part of a scientific community.

I also want to thank my high school Science Olympiad coach, Mr. Nicholas Stephens. Thanks to Mr. Stephens I found a home in science after moving from Puerto Rico to Washington. Mr. Stephens spend many after school hours allowing us to prepare for competitions and explore different scientific disciplines. I am lucky to have had such an amazing science high school teacher in path.

Throughout my journey in higher education, I have come across a community of supportive mentors. First, I want to acknowledge Lisa Peterson from the UW GenOM Project. There are not enough words I can say to express my gratitude to Lisa. Lisa has been an advocate for students of marginalized background in science. She has worked tirelessly to provide the tools for success at the UW for generations of students. I am lucky to have been part of the UW GenOM Project and to have been mentored by her. Next, I want to thank Theresa Britschgi my mentor from the Washington State Opportunity Scholarship. One of the most amazing parts of WSOS was having the opportunity to visit biotech industry sites, attend conferences and having Theresa as a mentor. I also want to thank Dr. Bruce Birren, Dr. Gisselle Velez and Francie Latour mentors from the Broad Summer Research Program who allowed me to feel part of an amazing scientific community at the Broad Institute. Finally, I want to thank my mentors at Ronald E. McNair Program: Dr. Todd Sperry, Quynh Tran, Dr. Cyndi Lopez, Dr. Monica Cortés Viharo. Thanks to the McNair Program, I had the tools to apply and succeed in graduate school. I am forever grateful for the support, encouragement, and advocacy mentors during my undergraduate years provided me with.

In 2019, I had to make the hard decision of which university to pursue my doctoral degree. One phone call was the deciding factor for me to join UCSF. I chatted on the phone with Dr. D'Anne Duncan who made it clear that if I decided to join UCSF, I would have the tools to succeed in graduate school. From that phone call I gained one of the most important mentors in my scientific journey. Dr. D'Anne Duncan is someone who I could rely on to give me honest feedback, she is someone who understood my strengths but also pushed me to work on my weaknesses. I also must thank Dr. D'Anne Duncan for all the unseen labor she does to create a space where UCSF students can thrive. I also thank Dr. D'Anne Duncan for the community she has created through IMSD and GRAD 210. Finally, I believe every graduate student deserves to have such a dedicated mentor as Dr. D'Anne Duncan and I am lucky to have her as my mentor.

I could not finish this acknowledgement section without thanking the people who I have been in this wild journey with. During my interview at UCSF in 2019, I had the amazing opportunity of meeting the people that would become my closest friends in graduate school: Oscar, Yash and Nebat. There are not enough words to explain how grateful I am to have found such supportive, talented, and trustworthy friends in my academic journey. Thank you for all the late nights we spent studying in first year, preparing for qualifying exams in second year, discussing research hurdles, and celebrating academic achievements. I also cannot forget to mention all the fun times we spent outside the lab and how you became my family in SF. Also, I want to thank other UCSF friends I made along the way specifically: Karissa, Matt, Jessica, Jackie, Sophie, Naz, Jackson, Jesslyn, Yanilka, Andres, Amaka, Chinaza, and Kingsly. It has been a tough journey, but I can't imagine not having y'all to go through it with.

I want to extend a thank you to the friends that supported me thorough graduate school. First, I want to thank my SF friends who provided me with much needed laughs outside of the lab Danny, Maurice, Dannytza, and Emerald. I want to thank my friend Karla who answered the phone many times after long lab days and gave me so much advice. Finally, I want to thank my friends, Rodha and Hilina, with whom I started this crazy journey at the UW GenOM Project. This amazing group of friends supported me through many difficult times and always made sure to uplift my spirits.

Finally, I want to thank my family. Thank you to the Laboy Reyes and Cintrón Carrasquillo families who have always reminded me where I come from and why I embarked on this journey. I want to thank my sister, Idalis, who is my best friend. I thank my sister for all her support throughout my life and graduate school. Most importantly, I want to acknowledge my parents, Isabel and David. This dissertation is dedicated to my parents. I could not have made it here without their sacrifices and hard work. I thank my parents for everything they have done to provide me with the best education. The work in this dissertation is possible due to the values and ethics my parents instilled in me from a young age. Words will never be able to encapsulate how thankful I am. I won the parents lottery, and I am so blessed to be able to have them as my parents.

CONTRIBUTIONS

Chapter 2 is adapted from an unpublished manuscript with the following authors: Michael Kosicki, Dianne Laboy Cintrón, Nicholas F. Page, Ilias Georgakopoulos-Soares, Jennifer A. Akiyama, Ingrid Plajzer-Frick, Catherine S. Novak, Momoe Kato, Riana D. Hunter, Kianna von Maydell, Sarah Barton¹, Patrick Godfrey, Erik Beckman, Stephan J. Sanders, Len A. Pennacchio, Nadav Ahituv.

Chapter 3 is adapted from an unpublished manuscript with the following authors:

Dianne Laboy Cintrón, Rory R. Sheng, Nadav Ahituv

FUNCTIONAL CHARACTERIZATION OF NEURONAL *CIS*-REGULATION

Dianne Laboy Cintrón

ABSTRACT

Most of the human genome does not encode proteins but instead contains a vast array of non-coding sequences that play crucial roles in gene regulation. Understanding these regulatory sequences, especially in neuronal contexts, is essential for understanding brain function and development. In this work, I utilized high-throughput assays alongside the mouse as a model to examine neuronal non-coding regulatory sequences. Chapter 1 provides a brief background of non-coding DNA. In Chapter 2, I tested thousands of candidate regulatory elements using Massively Parallel Reporter Assays (MPRA). We further validated strong candidates using mouse transgenic assays to assess the enhancer activity *in vivo*. Our combined approach of MPRA and mouse transgenic assays revealed complementary information on enhancer activity, highlighting the strengths and limitations of each method. In Chapter 3, I focused on functionally characterizing the regulatory network of the oxytocin receptor. The oxytocin receptor is a key regulator of social behavior. We identified seven candidate regulatory elements using comparative and functional genomics tools. We further validated the enhancer activity of the strongest candidate regulatory element using stable mouse transgenic lines. We determined the candidate regulatory element to have enhancer activity in the mouse olfactory bulb at post-developmental stages. This comprehensive study underscores the intricate regulatory landscapes that govern neuronal functions and showcases the power of integrating high-throughput screening with *in vivo* validation to unravel biological complexities.

TABLE OF CONTENTS

Chapter 1: INTRODUCTION.....	1
<i>Cis</i> -regulatory elements (CREs) in development and disease.....	2
Massively parallel reporter assays	3
Mouse transgenic assays.....	4
Non-coding variants association with psychiatric disorders	4
Characterization of neuronal <i>cis</i> -regulatory elements.....	5
References	6
Chapter 2: MASSIVELY PARALLEL REPORTER ASSAYS AND MOUSE TRANSGENIC ASSAYS PROVIDE CORRELATED AND COMPLEMENTARY INFORMATION ABOUT NEURONAL ENHANCER ACTIVITY.....	9
Abstract	9
Introduction	9
Results	12
MPRA neuronal library composition and initial QC.....	12
MPRA captures neuronal-specific activity	14
Neuronal MPRA activity correlates with mouse neuronal enhancer expression	18
Minimal MPRA effect of psychiatric disorder associated GWAS variants	19
Variants altering MPRA activity affect neuronal mouse enhancer activity.....	20
Discussion	21
Materials and Methods.....	42

MPRA design.....	42
LentiMPRA cloning and infection.....	44
Cell culture and neuronal differentiation.....	47
LentiMPRA analysis	48
Correlation of MPRA activity and epigenomic signal.....	48
TFBS enrichment analysis.....	49
Alignment and preprocessing of functional genomic data for ABC score pipeline .	49
Preprocessing of HiC and pHiC data for ABC score pipeline	50
Identification of candidate enhancer-gene pairs with ABC Score.....	51
Mouse enhancer transgenic assay.....	51
References	53
 CHAPTER 3: FUNCTIONAL CHARACTERIZATION OF <i>OXTR</i> -ASSOCIATED	
ENHANCERS.	61
Abstract	61
Introduction	61
Results	64
Annotation of <i>OXTR</i> cCREs.....	64
Luciferase assays of <i>OXTR</i> cCREs identify three functional enhancers.....	64
OCE7 is an active enhancer in the mouse olfactory bulb.....	66
Discussion.....	67
Materials and Methods.....	78
Candidate enhancer sequence selection	78

Luciferase Assays	78
Site-Directed Mutagenesis	79
Generation of transgenic mice.....	79
Quantitative RT-PCR.....	80
Immunostaining	81
References	82

LIST OF FIGURES

Figure 2.1: Functional validation of candidate cis-regulatory elements (cCREs) using lentiMPRA and mouse transgenic assays.	25
Figure 2.2: Neuronal WTC11 MPRA results validation.....	27
Figure 2.3: Predicting transgenic assay activity using a MPRA-based, coverage-corrected model.	29
Figure 2.4: Synthetic MPRA variants lead to in vivo change of function in transgenic assay.....	32
Supplementary Figure 2.1: MPRA quality control.	34
Supplementary Figure 2.2: Neuronal WTC11 MPRA results validation.	35
Supplementary Figure 2.3: Predicting transgenic assay activity using a MPRA-based, coverage-marginalized model.....	37
Supplementary Figure 2.4: Genomic tracks of the three nominally significant GWAS variants.	39
Supplementary Figure 2.5: Results of transgenic mouse assay.....	40
Figure 3.1: OXTR locus and OCE luciferase assays in hypothalamus cells.	69
Figure 3.2: OCE7 mouse enhancer transgenic assay shows enhancer activity in the mouse olfactory bulb at postnatal day 28.	70

LIST OF TABLES

Table 2.1: MPRA variants tested in transgenic mouse assay.	31
Supplementary Table 3.1: OCEs genomic coordinates and primers for cloning OCEs.	75
Supplementary Table 3.2: Linkage disequilibrium analysis of SNPs around OCE7.	76
Supplementary Table 3.3: Primer sequences.	77

Chapter 1:

INTRODUCTION

In the early 2000s, a monumental milestone in biology was achieved: the complete sequencing of the human genome¹. This achievement held the promise of solving many health challenges. However, scientists soon realized that much work remained to be done to understand the sequence that makes us human. Over the last 20 years, genetics research has focused on understanding the intricacies of the genome and how genetic changes can lead to human disease. In the journey to understand the genome, it quickly became clear that most DNA does not code for genes^{2,3}. This large portion of the DNA was initially labeled as "junk DNA," but it is more accurately described as non-coding DNA.

Following the Human Genome Project, individual labs and large research consortia focused on classifying the non-coding DNA into functional elements. The generation of publicly accessible data generated by projects like the Encyclopedia of DNA Elements (ENCODE) Project, the Functional Annotation of the Mammalian Genome (FANTOM) Project, Roadmap Epigenomics, Genomics of Gene Regulation Project, and the VISTA Enhancer Browser have been fundamental to guiding our understanding of the genome⁴⁻⁷. Other projects focused on capturing the full extent of human variation like the 1000 Genomes Project and Genome Aggregation Database (gnomAD) have been crucial to understanding how DNA changes can lead to disease⁸⁻¹⁰. Despite the abundance of resources, time, effort, and scientific ingenuity, much work is still needed to understand how non-coding regions of DNA function and to continue uncovering disease mechanisms that could lead to developing novel therapeutics.

Cis-regulatory elements (CREs) in development and disease

Most of our genome is non-coding, meaning it does not code for proteins. CREs are key components of non-coding DNA. These elements can be organized into distinct structures that regulate gene expression such as promoters, enhancers, insulators, and silencers¹¹. Promoters are located upstream of the gene and direct gene transcription through the binding of transcription machinery near the transcription start site^{6,11,12}. Insulators are bound only by CTCF and define boundaries of chromatin interactions¹³. Silencers are thought to repress the expression of the target gene through the competitive binding of transcription factors, harboring epigenomic marks associated with repression and binding of repressive transcription factors^{11,14,15}. Finally, enhancers dictate the transcription of a gene in a tissue-specific manner through binding specific transcription factors and interacting directly with gene promoters^{11,16}.

The functional importance of enhancers has been delineated through the disruption of key developmental enhancer elements of the β -globin gene (*HBB*) and sonic hedgehog (*SHH*), which lead to thalassemia and polydactyly, respectively¹⁷. CREs are crucial for maintaining gene regulatory networks and at the same time, most genetic variation across species and individuals of the same species occurs in these regions^{11,18}. Mutations or changes in CREs are thought to be the main drivers of *inter*- and *intra*-species phenotypic diversity¹⁸. However, unlike genes, which usually have a well-defined structure composed of 5' UTRs, promoters, exons, introns, poly(A) tails, and 3' UTRs, we still do not have a clear understanding of the grammar and structure of CREs. To this end, tools have been developed to identify and understand how CREs function and the impact of genetic changes in CREs on gene expression¹⁶.

Massively parallel reporter assays

Many tools have been developed to understand regulatory elements, specifically tools that biochemically annotate regions of DNA. First, Chromatin Immunoprecipitation Sequencing (ChIP-Seq) allows for the biochemical annotation of active enhancer marks, such as histone 3 lysine 27 acetylation (H3K27ac), and transcription factors. Second, DNase-Seq identifies open chromatin regions sensitive to DNase I digestion. Third, ATAC-Seq identifies open chromatin regions through the insertion of adapter sequences. These tools provide data on DNA regions that might be enriched in enhancer elements. Despite offering rich biochemical annotations, they do not provide functional readouts.

In this study, we employed Massively Parallel Reporter Assays (MPRAs), a technique that allows for the simultaneous characterization of thousands of regulatory elements¹⁹. In MPRA, a candidate enhancer sequence is cloned upstream of a minimal promoter, reporter gene, and barcode²⁰. Transcription of the barcode sequence allows for the quantification of the candidate enhancer element. MPRAs offer several advantages to other functional genomics tools. Unlike other tools mentioned above that only provide descriptive information, MPRA delivers a functional assessment. In 2020, the technology was further developed to use lentivirus for delivery, called lentivirus MPRA (lentiMPRA), which enables in-genome readout²⁰. Moreover, MPRAs can be performed *in vitro* across many cell types, allowing for the tissue-specific characterization of candidate CREs. Finally, MPRA provides a reproducible and quantifiable measure, allowing us to determine whether a sequence interrogated is active.

Like any technology, MPRA also has drawbacks. Specifically, the technology inquires into sequence activity in an out-of-locus context. Moreover, the length of tested sequences is limited by the available synthesizing technology. Despite these limitations, it remains a powerful tool for interrogating gene regulatory elements at scale.

Mouse transgenic assays

Before advancements in sequencing technologies and the development of high-throughput assays, scientists have used the mouse as a model to interrogate sequence activity *in vivo*⁷. These assays are labor-intensive, low to medium throughput, costly, and provide only qualitative measures. Despite these challenges, mouse transgenic assays have been used to characterize the spatiotemporal activity of enhancer elements *in vivo*. In general, mouse assays provide rich information on how enhancer elements might function in an organism.

Non-coding variants association with psychiatric disorders

In the past, genome-wide association studies (GWAS) have uncovered numerous non-coding variants linked to psychiatric disorders^{21,22}. The numbers of associated variants increase significantly since the lead single-nucleotide polymorphism (SNP) is not always the causative variant and there might be many other variants in linkage disequilibrium. Moreover, many of these variants are located within non-coding DNA, making it challenging to interpret their functional impact²¹. Individually testing each variant can be time-consuming and labor-intensive. Recent research studies have used MPRA to evaluate the effects of psychiatric disorder-associated variants at scale^{23–27}.

Characterization of neuronal *cis*-regulatory elements

MPRAs and mouse transgenic assays were used in the following chapters to characterize the regulatory activity of neuronal non-coding elements. In Chapter 2, we employed MPRA and mouse transgenic assays. Specifically, we tested the regulatory activity of thousands of elements via MPRA in differentiated neurons. We generated a catalog of functionally validated enhancer elements and characterized the impact of genomic variation in candidate CREs at scale. Finally, we determine that coupling MPRA and mouse transgenic assay data provides complimentary information on enhancer function.

Chapter 3 illustrates a hypothesis-driven approach to identify and characterize CREs of the oxytocin receptor (*OXTR*). *OXTR* is of particular interest as it plays a vital role in regulating behavior and disease²⁸. Much remains to be understood about how a single receptor can serve diverse roles. To address this question, I used a functional genomics approach. I investigated *OXTR* CREs which regulate the expression of *OXTR* across various tissues and lead to *intra*- and *inter*-species *OXTR* expression variation. Using publicly available data and data generated in our lab, we generated a list of candidate enhancers, evaluated their activity *in vitro*, and tested the strongest candidates *in vivo*. Using mouse models, we characterized the function of an enhancer element *in vivo*. We identified three active enhancers via luciferase assays and determined one to be an active enhancer in the olfactory bulb at postnatal days 28 and 56. Finally, we characterize the *OXTR* regulatory landscape and identify a novel olfactory bulb *OXTR*-associated enhancer.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Wong, G. K., Passey, D. A., Huang, Y., Yang, Z. & Yu, J. Is 'junk' DNA mostly intron DNA? *Genome Res.* **10**, 1672–1678 (2000).
3. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.* **6**, 699–708 (2005).
4. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
5. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
6. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
7. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–92 (2007).
8. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
9. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
10. gnomAD. <https://gnomad.broadinstitute.org/>.
11. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 45–63 (2017).

12. Riethoven, J.-J. M. Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. in *Computational Biology of Transcription Factor Binding* (ed. Ladunga, I.) 33–42 (Humana Press, Totowa, NJ, 2010).
13. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
14. Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F. & Orlando, V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat. Cell Biol.* **9**, 1167–1174 (2007).
15. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).
16. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
17. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
18. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2011).
19. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
20. Gordon, M. G. *et al.* Author Correction: lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* (2020) doi:10.1038/s41596-020-00422-z.
21. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol.*

- Genet.* **24**, R102–10 (2015).
22. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu & Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469–1482.e11 (2019).
 23. Deng, C. *et al.* Massively parallel characterization of regulatory elements in the developing human cortex. *Science* **384**, eadh0559 (2024).
 24. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).
 25. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).
 26. Rummel, C. K. *et al.* Massively parallel functional dissection of schizophrenia-associated noncoding genetic variants. *Cell* **186**, 5165–5182.e33 (2023).
 27. Zeng, B. *et al.* Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases. *bioRxiv* (2023)
doi:10.1101/2023.03.02.530826.
 28. Jurek, B. & Neumann, I. D. The Oxytocin Receptor: From Intracellular Signaling to Behavior. *Physiol. Rev.* **98**, 1805–1908 (2018).

Chapter 2:

MASSIVELY PARALLEL REPORTER ASSAYS AND MOUSE TRANSGENIC ASSAYS PROVIDE CORRELATED AND COMPLEMENTARY INFORMATION ABOUT NEURONAL ENHANCER ACTIVITY.

Abstract

Genetic studies find hundreds of thousands of noncoding variants associated with psychiatric disorders. Massively parallel reporter assays (MPRAs) and *in vivo* transgenic mouse assays can be used to assay the impact of these variants. However, the relevance of MPRAs to *in vivo* function is unknown and transgenic assays suffer from low throughput. Here, we studied the utility of combining the two assays to study the impact of non-coding variants. We carried out an MPRA on over 50,000 sequences derived from enhancers validated in transgenic mouse assays and from multiple fetal neuronal ATAC-seq datasets. We also tested over 20,000 variants, including synthetic mutations in highly active neuronal enhancers and 177 common variants associated with psychiatric disorders. We found a strong and specific correlation between MPRA and mouse neuronal enhancer activity. Four out of six tested variants with nominally significant MPRA effects affected neuronal enhancer activity in mouse embryos. Mouse assays also revealed pleiotropic variant effects that could not be observed in MPRA. Our work provides a large catalog of functional neuronal enhancers and variant effects and highlights the effectiveness of combining MPRAs and mouse transgenic assays.

Introduction

Genome-wide association studies (GWAS) have identified hundreds of non-coding variants associated with psychiatric disorders, which exhibit complex genetic

etiologies likely involving multiple loci¹⁻⁶. The GWAS-discovered lead variants are not necessarily causative due to linkage disequilibrium (LD), which increases the number of potential variant candidates on average by ten-fold. In addition, ongoing whole genome sequencing studies of patients with psychiatric disorders identify approximately 70 *de novo* non-coding variants per individual⁷. These efforts highlight the challenge to distinguish causative variants from the hundreds of thousands of potential candidates identified through genetic studies.

Various genomic correlates of function can be used to reduce the number of potential candidates. Putative regulatory sequences can be identified in a tissue and even cell-type specific manner using such methods as DNase-seq and ATAC-Seq (for identification of open chromatin), or ChIP-seq⁸⁻¹² (for identification of regions bound by transcription factors or having specific histone marks). Variants falling into regulatory regions with activity in relevant cell types are more likely to be causative. However, an overlap between a variant and regulatory region neither confirms variant functionality, nor provides a mechanism for how it impacts the phenotype. Functional assays that can test the effect of the variant on gene regulatory activity are needed to pinpoint the causative mutations.

Massively parallel reporter assays (MPRA) allow for the assessment of regulatory activity of tens of thousands to hundreds of thousands of candidate regulatory sequences and variants within them¹³⁻¹⁵. The majority of MPRA are conducted *in vitro*, enabling the high throughput interrogation of candidate sequences and variants in a quantitative and reproducible manner. However, they are limited to testing the function of the assayed sequence only in the specific cell type and cannot

assess how results relate to its function *in vivo*. As an alternative, *in vivo* activity of enhancers can be tested using a transgenic mouse assay (referred to as "transgenic assay" below) such as enSERT^{16,17}. In this assay, a candidate regulatory sequence is coupled to a minimal promoter and reporter gene followed by its integration into a safe harbor locus in mouse zygotes and assayed for activity by imaging at a later embryonic time point. Transgenic assays can identify enhancer expression at an organismal level, providing rich, multi-tissue phenotype. Results of thousands of these assays are cataloged in the VISTA enhancer browser and serve as a gold standard for enhancer activity assessment¹⁸. However, these assays are more resource and labor intensive than MPRA and therefore are typically conducted at a much lower throughput. Combining the high throughput capabilities of MPRA and rich phenotype of transgenic assays is an underexplored venue for regulatory element and variant characterization. Limited comparisons of these technologies have been performed^{19–23}, but typically involved MPRA conducted in cancer or immortalized cell lines with limited relevance to organismal biology, used short sequences (120 bp) or sampled too sparsely from *in vivo* validated sequences to enable a systematic comparison.

Here, we set out to robustly compare results between MPRA and transgenic assays by using psychiatric disorders-associated sequences and variants as a test case. We carried out an MPRA for over 50,000 sequences 270 bp in length, many of which were derived from brain enhancers in the VISTA enhancer browser¹⁸ and over 20,000 variants. We found thousands of functional regulatory sequences and hundreds of variants that alter regulatory activity compared to their reference allele. We observed an overall strong correlation between MPRA and transgenic assays. Variants with a

high impact in MPRA also had a significant effect on neuronal enhancer activity in transgenic assays in mouse embryos. Combined, our work provides a large catalog of functional neuronal enhancers and their variants and shows that MPRA can be successfully combined with mouse transgenic enhancer assays.

Results

MPRA neuronal library composition and initial QC

We set out to investigate the correlation between high-throughput MPRA and mouse enhancer transgenic assays. As neuronal enhancers are the most abundant category of enhancers in the VISTA Enhancer Browser¹⁸, which catalogs mouse transgenic assay results, and since our lab has established MPRA protocols in stem cell differentiated neurons^{19,24,25}, we focused on neuronal-associated elements. We designed an MPRA library by tiling peaks from five single-cell and bulk neuronal ATAC-seq experiments^{26–30} and from conserved cores of 1,400 neuronal and non-neuronal enhancers from the VISTA Enhancer Browser¹⁸ with 270 bp tiles (**Figure 2.1a,b**; minimum 30 bp overlap; see Methods). To characterize how mutations affect the activity of these elements, we introduced two types of variants into the designed tiles. First, we included all lead single-nucleotide polymorphisms (SNPs) and SNPs in linkage disequilibrium ($r^2 > 0.8$) from autism spectrum disorders (ASD), schizophrenia, bipolar disorders and depression GWAS that overlapped designed tiles^{1,3–5}. Second, we introduced synthetic transversion variants into every fourth base pair of elements with high likelihood of MPRA activity (overlapping ATAC-seq peaks from multiple datasets, evolutionary conserved, active in transgenic assay, see Methods; **Figure 2.1a,b**)³¹. As negative controls, we used 500 di-nucleotide scrambled, non-conserved tiles from

enhancers negative in mouse transgenic assays that did not have overlapping ENCODE candidate *cis*-regulatory elements³² or neuronal ATAC-seq signal^{26–30}. In total, we designed 81,952 unique 270 bp sequences, including 24,942 variants.

Oligos were synthesized and cloned into a barcoded lentiMPRA vector and packaged into lentivirus following our previously published protocol¹⁵. They were then transduced into differentiated human excitatory neurons derived from an isogenic WTC11-Ngn2 iPSC line with an inducible Neurogenin-2 gene using an established induction protocol^{15,34,35}. Only tiles with at least 15 barcodes detected in each of the three replicates were retained (median = 177 barcodes post-filtering; **Supplementary Figure 2.1a**) and tiles with mutations without a reference tile passing these criteria were discarded. Out of 81,952 elements, 76,415 passed QC (> 90%; see Methods), including 52,335 genomic elements, 23,482 single base pair mutation tiles and 476 scramble negative controls. Together, the elements covered 11.7 Mbp of genomic sequence in 24'000 non-overlapping regions of 270 bp (tile size) to 6531 bp in size (mean 488 bp). MPRA activity was expressed as a z-score of $\log_2(\text{RNA counts}/\text{DNA counts})$ relative to scramble negative controls. Negative control reference tiles, which were selected from non-conserved parts of elements negative in transgenic assay and with no epigenomic signal in neural datasets, had a similar activity to their scrambled counterparts (**Supplementary Figure 2.1b,c**). This both validated their selection strategy and showed that scrambling did not systematically make elements active or repressive. We observed good correlation between replicates (Pearson correlation = 0.58-0.59, **Supplementary Figure 2.1b**, N = 76,415). Based on nominal p-value < 0.05 from a t-test against mean of scramble negatives and an absolute activity at least one

standard deviation above scrambled controls, we designated 4,762 tiles to be activators and 2,957 tiles to be repressors (out of 52,335, 14.7%). Using similar criteria, we found 467 single base pair mutation tiles to have decreased activity compared to reference tile and 313 to have an increased one (out of 23,482, 3.3%). These designations should be treated as operational since none of the tiles or variants remained significant after multiple testing correction.

MPRA captures neuronal-specific activity

To validate the results of our MPRA, we annotated the activity of tiles overlapping a variety of genomic annotations. Specifically, we asked if ranks of the overlapping tiles were significantly different than scrambled negative controls (**Figure 2.2a**; **Supplementary Table 2.1**). On average, elements in our library were more active than scrambled negative controls (median activity = 0.19). Overlap with positive elements in previous neuronal MPRA was associated with higher activity, with elements from Inoue 2019¹⁹ publication (double-Smad inhibition protocol) being more active than those from Uebbing 2021³⁶ (stable neural stem cell line, median activity 0.31 vs 0.18). We also confirmed that tiles that were pre-selected for mutagenesis due to high expected activity were indeed highly active ("Mutation reference tiles", activity = 0.33). At the positive extreme, tiles overlapping housekeeping promoters (defined as 2 kb centered around the 5' end of Gencode protein-coding exon 1 of genes in Eisenberg and Levanon 2013³⁷) were highly active (median activity 0.56), suggesting that they can be used as universal positive controls in MPRA. We note that they may function as autonomous promoters and not as enhancers³⁸. Ultraconserved elements³⁹ had high activity as well (0.41), which is consistent with our previous observation that they are often active in

developing brain in the transgenic assay⁴⁰. Conversely, tiles overlapping coding exons, but not exons of long-non-coding RNAs, were overall repressive (median activity = -0.23). It is unlikely splicing sites present at exon-intron interface explain this observation, as in our MPRA the minimal promoter, and consequently the transcription start site, are downstream of the tested element. In addition to rank-based analysis, we also checked if the fraction of activator and repressor tiles overlapping a given genomic annotation was significantly different from that of tiles that did not overlap it (i.e. we split all tiles into four groups based on activator/repressor status and overlap with a given annotation, calculated log(odds ratio) and performed a Fisher exact test). The results obtained using this method were similar (**Supplementary Figure 2.2a; Supplementary Table 2.1**).

We then set out to analyze the transcription factor binding sites (TFBS) to further validate our MPRA captures neuron-specific signals. Using HOMER⁴¹, we compared activator or repressor tiles that do not overlap promoters (N=3,054 and 1,894) to either genomic background or to other tiles from our MPRA with background level activity ("scramble-like"; $-0.4 < \text{activity} < 0.4$, N = 15,503; **Figure 2.2b; Supplementary Table 2.2**). We used HOMER mouse and human TFBSs (N = 439). The analysis accounts for GC-content differences in test and background sets. We considered a TFBS to be enriched if it was present in at least 10% test tiles, increased by at least 50% compared to background set (corresponding to $\log_2(1.5) = 0.58$ cutoff) and was significantly enriched by HOMER's hypergeometric test at FDR-adjusted p-value < 0.01. We found a total of 13 motifs to be enriched in activator tiles compared to tiles with background activity, including neuron-associated motifs SP5, KLF1, CUX2 and five motifs from the

RFX family^{42,43}, as well as a NFY-binding promoter motif CCAAT, two liver, pancreas and nervous system expressed HNF6/ONECUT1 motifs⁴⁴, pituitary-development associated PIT1/POU1F1 (POU family) and growth/survival TF ATF1. This demonstrates our MPRA captured some neuron-specific signal. Analogous analysis using genomic background found the same motifs (except ATF1 and KLF1) and a score of additional, neuronal-associated ones, mostly from SOX, LHX, DLX and E-box families (NEUROD1, MYOD, ATOH1; **Supplementary Figure 2.2b,c,d**)^{43,45}. Repressor tiles were enriched for similar motifs as activators when compared to genomic background. In particular, we observed SOX, LHX and DLX families, with only repressor-specific hits being SOX3 and BORIS/CTCF (**Supplementary Figure 2.2b,d**). No repressor-enriched motifs were found when comparing to tiles with background-level activity. This may imply lack of specific repressive signal in our library, limited power to detect such signal or simply reflect a relative dearth of known repressive motifs in HOMER dataset. The latter might be consistent with de novo motif analysis conducted against tiles with background activity, which revealed similar motifs for activator tiles (RFX and CUX2, match scores 0.94-0.96 out of 1), but novel repressor motifs with tentative matches to among others ETV4, PAX5, ZNF135, MYOD1 and ZEB1 (match scores of 0.65-0.72; **Supplementary Table 3**).

We observed that both activator and repressor tiles had higher median levels of GC-content than the rest of the library, with repressors having higher levels than activators (repressors 64%, activators 50%, remaining elements 44%; **Supplementary Figure 2.2e**). Such GC-skew should not affect MPRA readout on a technical level, as the activity of the tested element is read through sequencing of a barcode, not the

element itself (unlike e.g. in STARR-seq). We conclude that highly GC-rich sequences may function as transcriptional repressor elements in this MPRA.

We next set out to assess how well various biochemical marks correlated with neuronal MPRA activity. To investigate whether activity signal in our MPRA is biologically specific, we compared our MPRA results to epigenomic signal from diverse tissues and cell types from 12 embryonic, fetal and WTC11 datasets, encompassing 740 Dnase hypersensitive sites (DHS), ATAC and single-cell ATAC samples (**Supplementary Table 4**). To account for a large diversity of experimental and computational protocols, we integrated raw genomic signal (bigWig tracks) over MPRA tiles and ranked the tiles based on the signal for each dataset. We then computed the difference between median MPRA activity of top ranked elements and the remaining ones for a range of epigenomic rank cutoffs (**Figure 2.2c**). As expected, the more stringent the rank cutoff, the larger the difference between activity of top ranked elements versus the rest. However, due to enrichment of promoter-overlapping elements in top ranks, the differences between individual datasets was negligible (**Figure 2.2d, left**). After removing tiles overlapping protein-coding promoters, we observed a clear separation of brain and differentiated WTC11 cells samples from non-neuronal samples (**Figure 2.2d, right**). Closer inspection revealed that some non-neuronal ENCODE DHS samples (adrenal, eye and kidney) are still enriched, especially at stringent cutoffs, possibly reflecting a combination of high activity tissue-invariant elements ("housekeeping enhancers") and higher signal-to-noise ratio of DHS data at high signal intensities. This was attenuated at less stringent signal cutoffs, with only four non-neuronal samples (eye and adrenal) remaining in top 50 at signal rank cutoff 5000

(**Supplementary Table 2.4**). Encouragingly, we observed a clear separation over ATAC-seq time course of WTC11 cell neuronal differentiation, with undifferentiated cells ranking at position 599, day 3 differentiated cells at position 62 and day 14 at position 7⁴⁶. We note that our MPRA design sampled elements with open chromatin signal in neuronal tissues more deeply than in other tissues, which may have contributed to the observed enrichments. In summary, our results show that our MPRA captures neuronal-associated regulatory activity.

Neuronal MPRA activity correlates with mouse neuronal enhancer expression

The average sequence length tested in transgenic mouse assays is around 1 kb, about four times the size of tiles in our MPRA (270 bp). To compare these two assays, we matched transgenic assay elements ("VISTA elements") with overlapping MPRA tiles of highest activity (**Figure 2.3a**). We then built a general linear model (GLM) with a binomial link to predict binary, tissue-specific transgenic assay results (e.g. brain activity, yes or no) from MPRA activity. In our design, we have included negative control tiles derived from non-conserved parts of negative VISTA elements that did not overlap epigenomic signal from any of the neural datasets. Conversely, we aimed to capture as many conserved parts of neural-positive VISTA elements as possible (**Figure 2.3a**). To account for that design bias, we included a fraction of conserved sequences covered by tiles as a covariate in the model (**Figure 2.3b,c**). An alternative solution, removing poorly covered VISTA elements yielded similar results (**Supplementary Figure 2.3a,b,c**). A model without any filtering and covariates is included for completeness (**Supplementary Figure 2.3d**).

We found that all neural annotations (except dorsal root ganglion) were significantly correlated with MPRA activity, while craniofacial and heart terms were significantly negatively correlated (**Figure 2.3c**). The steepest regression slope on the positive side was for a combined 'neural' term (brain, neural tube, cranial nerves including trigeminal nerve and dorsal root ganglion), followed by 'brain'. We also validated this approach using an independent, published MPRA conducted in primary human fetal cortical cells⁴⁷. Forebrain and combined brain terms were positively correlated with MPRA activity, while heart, heart+somite and facial-mesenchyme were negatively correlated, similar to our MPRA (**Supplementary Figure 2.3e**). We note that much fewer VISTA elements were overlapped by tiles in this MPRA (386 vs 1400), which likely accounts for fewer significant terms. Further, while our MPRA took conservation into account during design, Deng 2024 MPRA was free of this assumption, which explains why conservation-coverage regressed model and model without conservation coverage covariate behaved nearly identically (**Supplementary Figure 2.3e**). We conclude that neural MPRA in differentiated human excitatory neurons and neural activity in transgenic assay strongly correlate in a tissue-specific manner.

Minimal MPRA effect of psychiatric disorder associated GWAS variants

We next analyzed the 177 psychiatric disorder associated GWAS variants tested in our MPRA. Using nominal significance criteria (see Methods), we found that only 3 out of the 177 variants had a significant effect on MPRA activity (**Supplementary Table 2.5**). Each variant was associated with an independent GWAS signal (different lead SNPs, two associated with bipolar disorder, one with major depressive disorder) and had a moderate, gain-of-activity impact on expression (1.1-1.5 scrambled negative

standard deviation units). None of the variants remained significant after multiple testing adjustment, all were in tiles with either repressive or no activity (-1.79 to 0.13) and tended to be outside or at the edge of the peak of the DHS signal and conservation (**Supplementary Figure 4, Supplementary Table 2.5**). Therefore, we decided not to further investigate them using the transgenic assay.

Variants altering MPRA activity affect neuronal mouse enhancer activity

To select variants for transgenic assay follow-up, we analyzed the synthetic, single nucleotide variants. Out of 23,266 tiles with non-GWAS variants, 777 had a nominally significant effect on regulatory activity (p-value < 0.05, absolute log₂ fold-change > 1). We note that none of the variants remained significant after multiple testing correction. We selected six of these variants for follow up in the transgenic assay, based on prior evidence of neuronal activity in transgenic assays (5/6 variants) and, to a lesser degree, links to important neuronal genes predicted using ABC⁴⁸ (e.g. *QKI*, *PRKN*, *COA7*, *SETBP1* and *MEF2C*; **Table 2.1; Figure 2.4a**).

We found that 4/6 variants affected mouse enhancer expression in a reproducible manner, with 4 causing a loss of activity in different parts of the brain, neural tube or cranial nerves (**Figure 2.4b**). In two cases, this was accompanied by a gain of expression in another brain-associated structure. We note that the two variants with no apparent impact on transgenic enhancer activity had a very high basal activity of the reference element in the transgenic assay (hs268), which may have masked expression differences due to the variant. These results demonstrate the utility of combining the two experimental systems, with a good correspondence between MPRA and mouse transgenic assay and rich additional information provided by the latter.

To further interpret the results of our transgenic assays, we used motifbreakR⁴⁹ to carry out TFBS predictions for all six variants tested using transgenic assay. We found plausible hits for all variants except hs268.2, which did not have an effect in the transgenic assay (**Figure 2.4c,d**). In four cases, more than one plausible TFBS was found. We leveraged the fact our MPRA design also contained variants in close proximity to the ones we selected for transgenic assay follow up to further validate the TFBS predictions. For example, the variant tested in hs978.1 element was predicted to both create a potential repressor CDX1 site⁵⁰ and destroy the POU4F3 site. The flanking variant MPRA effects was more consistent with CDX1 creation i.e. the flanking variant that did not affect MPRA activity was also predicted not to affect CDX1 binding (**Figure 2.4c**). Conversely, POU4F3 binding was not consistent with two of the flanking variants. Both of these variants had high predicted relevance for POU4F3 binding (based on POU4F3 motif), but exerted no effect on MPRA activity (**Figure 2.4c**). We applied similar logic to remaining predictions to select the more plausible of the initial TFBS matches. Deploying this approach in a systematic manner could help interpret future variant MPRA.

Discussion

We performed an MPRA in neurons with elements derived from VISTA enhancers and neuronal fetal ATAC-seq peaks finding a good correlation to neuronal expression in mouse transgenic assays. In terms of variants, we did not see a strong effect on MPRA activity for our selected psychiatric disorder associated GWAS variants, but observed effects on MPRA activity for 777 out of 23,266 synthetic variant tiles. Four out of six synthetic variants nominated by MPRA as having a nominally significant

impact also affected the transgenic assay activity in the expected manner and revealed additional ectopic effects. Overall, we demonstrated that combining MPRA and transgenic assays can be highly advantageous.

The observed complementary of the two assays is encouraging. MPRA allows the testing of a large number of sequences and provides a quantitative readout, while transgenic assays can reveal the organismal spatially-resolved consequences of regulatory sequences and variants. Both approaches have improved significantly over the past decade, coming closer to bridging the gap between them. MPRA has been increasing in throughput, length of tested elements, range of cell types amenable to this type of assay (due to lentivirus and AAV delivery) and have also been carried out *in vivo* in select tissues in a postnatal manner^{15,22,23,51–53}. Transgenic assays improved in throughput and reproducibility due to the development of Cas9-guided safe harbor integration method enSERT¹⁷. Shortcomings of MPRA and transgenic assays, as listed below, can be overcome by combining these techniques. MPRA conducted *in vitro* are limited to the cell types in which they are assayed, can be limited by the availability of differentiation protocols and labor intensity of differentiating millions of cells with various identities and cannot assess the spatial and temporal organismal activity of a regulatory element. enSERT is conducted in mice, which cannot capture all aspects of human biology, is costly and not high-throughput. It also has only recently been applied in a quantitative manner⁵⁴. While both methods are likely to improve and eventually merge, our work highlights the utility of combining currently available approaches, with MPRA as a high-throughput filter for the multi-tissue transgenic assay. We note this finding

may not fully translate to episomal MPRA, whose results may be less reliable than those of lentiMPRA used in this study.

We did not observe a significant effect on MPRA activity for the 177 tested psychiatric disorder associated GWAS variants after applying multiple testing correction. This is in line with another MPRA carried out by our lab that found only 164 psychiatric disorder and eQTL variants out of 17,069 tested (< 1%) to have an effect on MPRA activity⁵³. This could be due to a variety of reasons: 1) The small number of variants tested and low statistical power; 2) Generally low expected effect size of common variation associated with complex traits such as psychiatric disorders. Machine learning models of MPRA data⁵³ suggest that rare variants have a higher effect on MPRA activity compared to common variants; 3) Some variants may affect non-transcriptional phenotypes, like chromatin tethering⁵⁵; 4) Some variants may have an effect in another cell type or at a different differentiation time point.

Synthetic variants comprised most variants tested in this MPRA, which has some advantages over testing common variants. First, the effect sizes of these variants are not constrained by negative selection, unlike common variation in human populations. This makes synthetic MPRA a better substrate for computational modeling, which should be able to learn a wide range of potential effects. Second, our experiment allowed us to find functional variants in elements likely to control expression of neuronal genes, some of which are linked to neurodevelopmental disorders. These results place a strong prior on interpretation of yet undiscovered, large effect de novo variants in these regions and can help better understand the regulatory biology of neuronal development.

In summary, we compiled a catalog of transcriptional activity in neuronal cells of over 50,000 elements derived from open chromatin fetal datasets and enhancers validated in transgenic assays. We also assessed the impact of over 20,000 synthetic and 177 GWAS variants and demonstrated the usefulness of using MPRA as a variant filter for transgenic mouse assays. We anticipate this work will contribute to computational modeling of gene regulation and studies focused on neural development and psychiatric disorders.

Figures and Tables

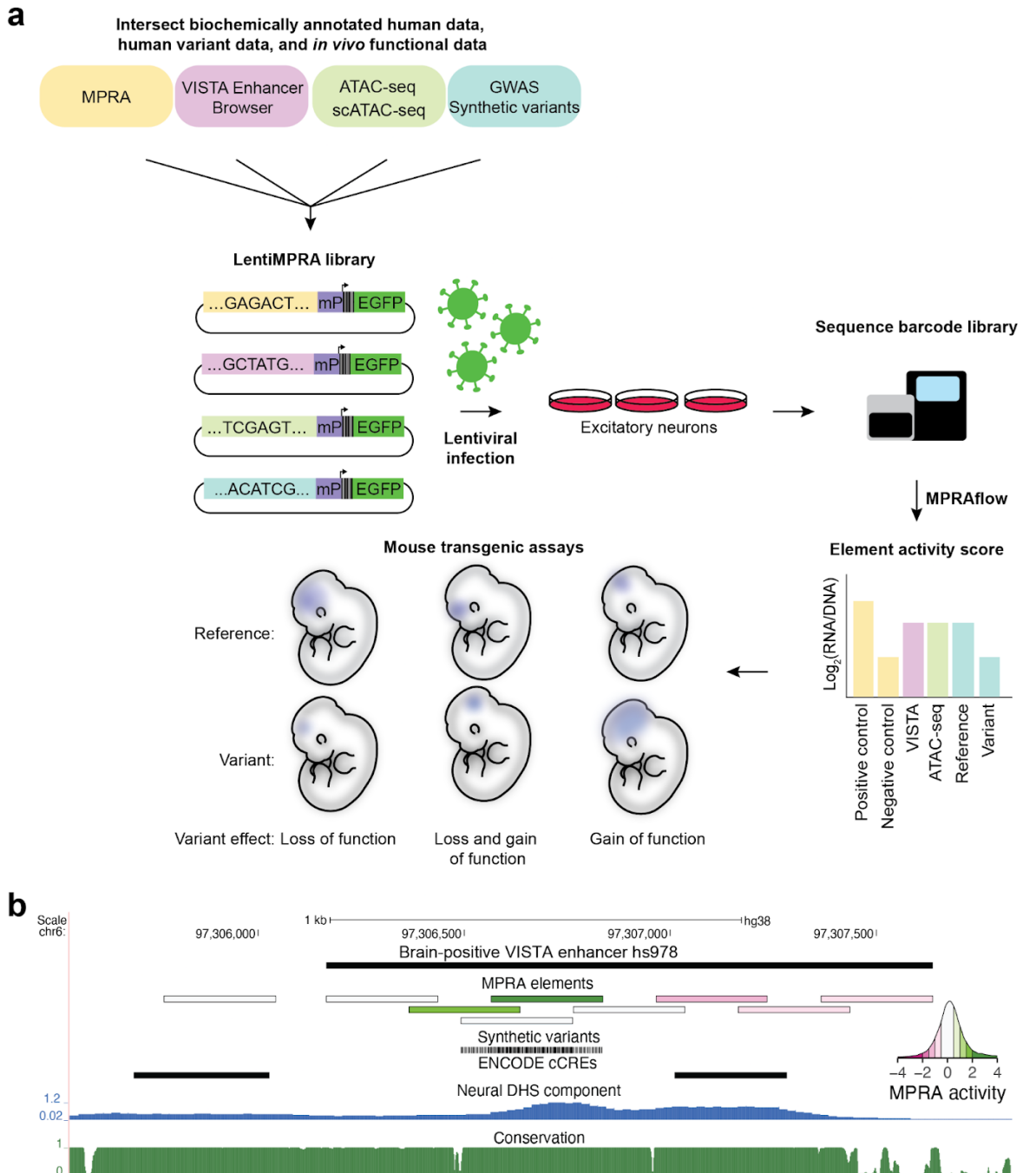


Figure 2.1: Functional validation of candidate *cis*-regulatory elements (cCREs) using lentiMPRA and mouse transgenic assays.

(a) Schematic of experimental plan. A lentiMPRA library was designed through the intersection of scATAC-seq, ATAC-seq, VISTA Enhancer Browser¹⁷, conservation and neuronal MPRA data. The library also included GWAS lead SNPs and SNPs in LD with them and synthetic variants. Sequences were inserted into a reporter plasmid upstream (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
of a minimal promoter (mP), barcode and EGFP. The library was infected into WTC11 induced excitatory neurons using lentivirus. The integrated DNA and transcribed RNA barcodes were sequenced to determine element activity scores. Mouse transgenic assays were conducted on selected sequences to characterize their activity *in vivo*. **(b)** UCSC Browser annotation, from top to bottom: (1) VISTA enhancer browser hs978 sequence (2) MPRA elements colored by MPRA activity with green showing high activity and pink low activity (see inset). (3) synthetic variants included in MPRA tested elements (4) ENCODE cCRE (candidate *cis*-regulatory elements)³¹ (5) Neural DNase I hypersensitivity signal component³² (6) PhastCons conservation UCSC track for 30 mammals (27 primates).

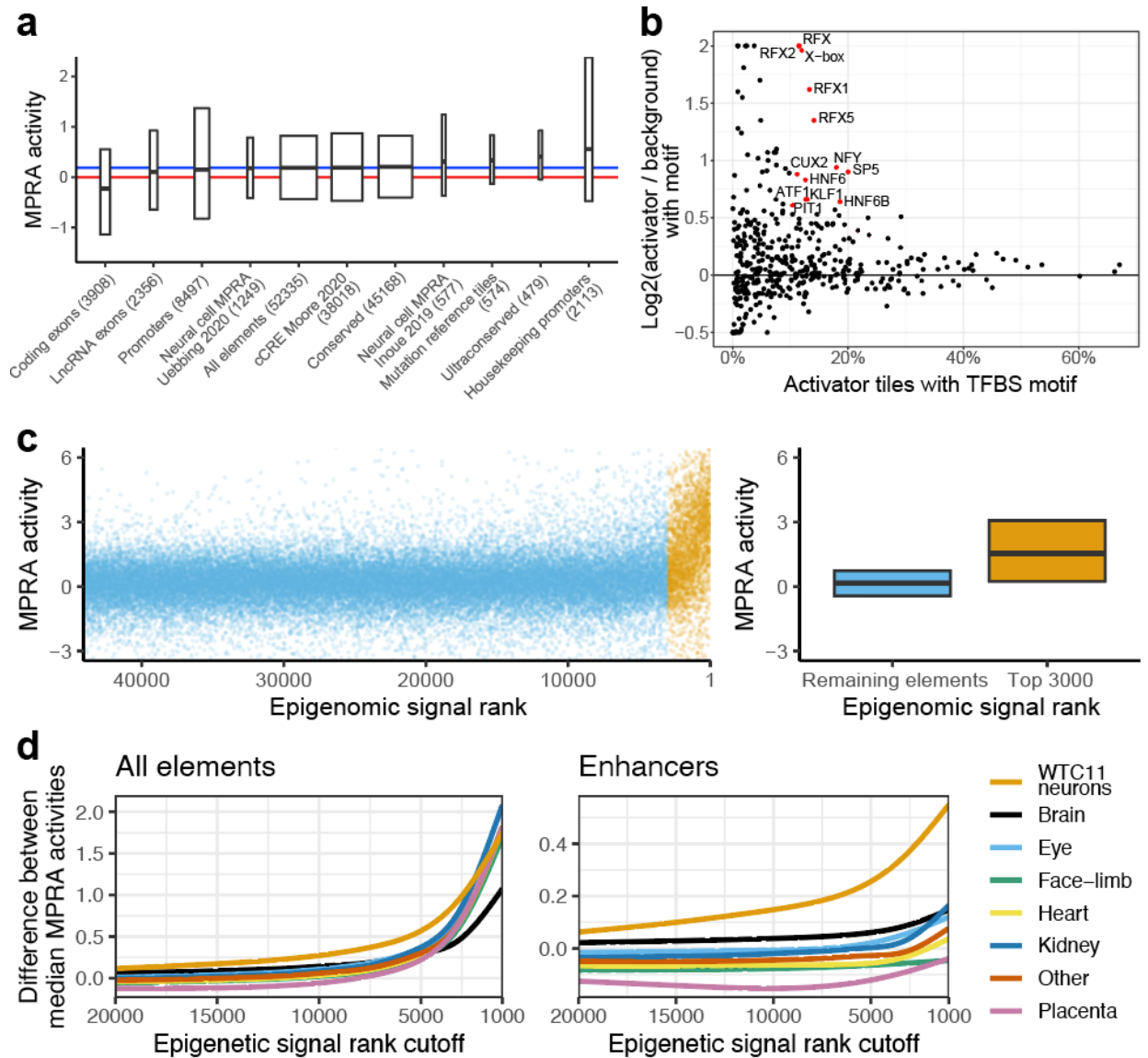


Figure 2.2: Neuronal WTC11 MPRA results validation.

(a) MPRA activity of tiles overlapping different categories. Red line = activity of scrambled negative controls (zero, by definition). Blue line = median activity of all reference elements (0.19). Hinges of boxplot span interquartile range, line in the middle is median, width is proportional to the number of overlapping tiles. All categories have significantly different activity than scrambled negatives at FDR-adjusted p-value < 0.05 (Mann-Whitney U test). Promoters = 5' end of exon 1 of protein-coding genes +/- 1 kb.

(b) TFBS enrichment in enhancer, activator tiles compared to enhancer elements with scramble negative levels of activity. Log₂-fold change was curbed at -0.5 and 2. Only TFBSs present in more than 10% target, with 50% increase in presence from background to target set (corresponding to log₂(1.5) = 0.58 cutoff) and FDR < 1% are labeled and colored red.

(c) Methodology for comparison of epigenomic annotation. Left: tiles were ranked by epigenomic signal and split at various rank cutoffs into two (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
groups. Right: Median MPRA activity of the two groups was compared. Same boxplot display conventions as in (a). **(d)** Difference in median MPRA activity at different epigenetic rank cutoffs for eight tissue groups. Left: all elements (N = 44,109; lower than 52,335 total due to exclusion of elements that failed to lift-over between human and mouse genomic assemblies), right: enhancers (N = 37,813; enhancers defined as not overlapping "coding promoter" category in (a); see Methods).

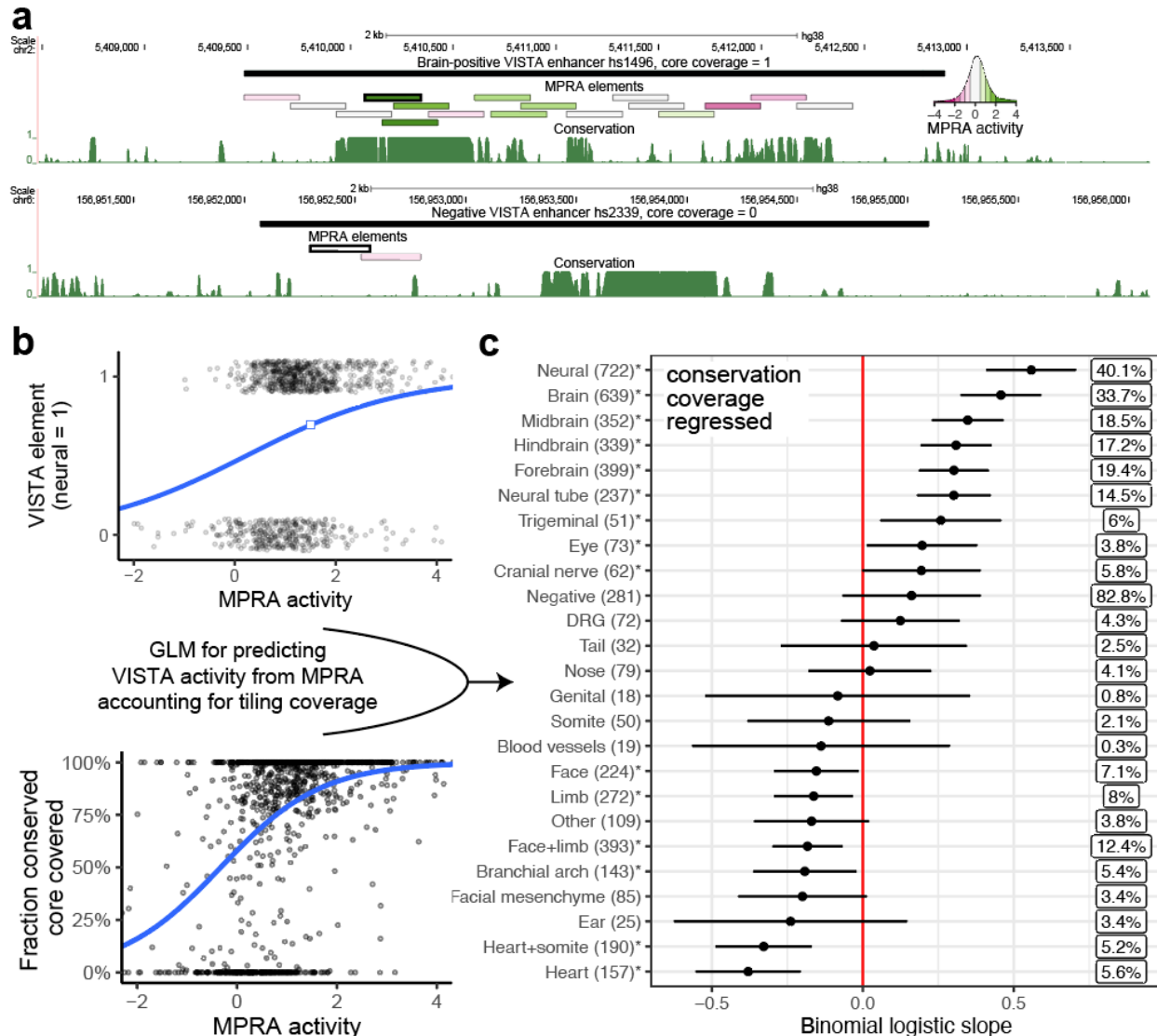


Figure 2.3: Predicting transgenic assay activity using a MPRA-based, coverage-corrected model.

(a) Examples of VISTA elements with complete (top) and zero (bottom) coverage of their conserved cores using MPRA tiles. Conservation is PhastCons UCSC tracks for 30 mammals (27 primates) dataset. MPRA tile with highest activity (used for modeling) has a thicker border. MPRA elements colored by MPRA activity, see inset. (b) Visualization of input variables for the GLM. Top: transgenic assay (VISTA) elements are binarized according to chosen tissue activity (here: neural; jitter added for visualization). The blue line is the binomial-link GLM regression on this variable. Bottom: relationship between fraction of conserved core covered and MPRA activity is modeled as a covariate. The blue line is the binomial-link GLM regression on this variable. (c) Results of the GLM predicting binomial transgenic assay activity from MPRA activity and fraction of conserved core covered. Asterisks indicate nominal p-value < 0.05. Boxed percentages to the right are Nagelkerke R^2 measures. Bars extend two standard errors of the mean (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
in each direction. DRG = dorsal root ganglia. Face-mesen = facial mesenchyme. Cranial
nerves category does not include the trigeminal nerve, as per VISTA Browser.

Table 2.1: MPRA variants tested in transgenic mouse assay.

TFBS predictions from motifbreakR. Motifs consistent with flanking variant effects in bold (see **Figure 2.4c**). Neural target genes found using activity-by-contact (ABC) on data from WTC11 excitatory neurons (cell line used in this study) or prenatal week 18 prefrontal cortex neurons (Methods). Asterisk - element not previously tested in the transgenic assay.

Variant name	Reference MPRA activity	Variant MPRA effect	TFBS affected	VISTA element	Structures affected in transgenic assay	Neural target genes	Relevant target gene Phenotypic associations
chr6-97306759-A-T	4.1	-4.5	CDX1 (gain), POU4F1	hs978.1	partial forebrain, hindbrain, and neural tube loss	GPR63	Branchiooculofacial Syndrome and Spina Bifida Occulta
chr6-162856979-C-G	3.8	-4	RORB (gain), CTCF	hs2793.1	partial forebrain loss	QKI, PACRG and PRKN	Associated with Parkinson's Disease (PACRG, PRKN) and schizophrenia (QKI)
chr1-52663061-C-G	6.2	-4.5	SP2	hs2790.1*	midbrain gain, partial forebrain loss	COA7, TUT4 and others	Spinocerebellar Ataxia (COA7) Perlman Syndrome (TUT4)
chr18-44826694-C-G	0.6	-2.6	MAFB	hs2791.1	partial cranial nerve loss, midbrain, hindbrain and neural tube gain	SETBP1	Autism
chr5-88396638-C-G	-0.3	-2.5	ASCL1 (gain), NR2F1	hs268.1	no effect	MEF2C	Associated with cognitive disability, epilepsy and cerebral malformations
chr5-88397333-A-T	0.6	-1.5	[none]	hs268.2	no effect		

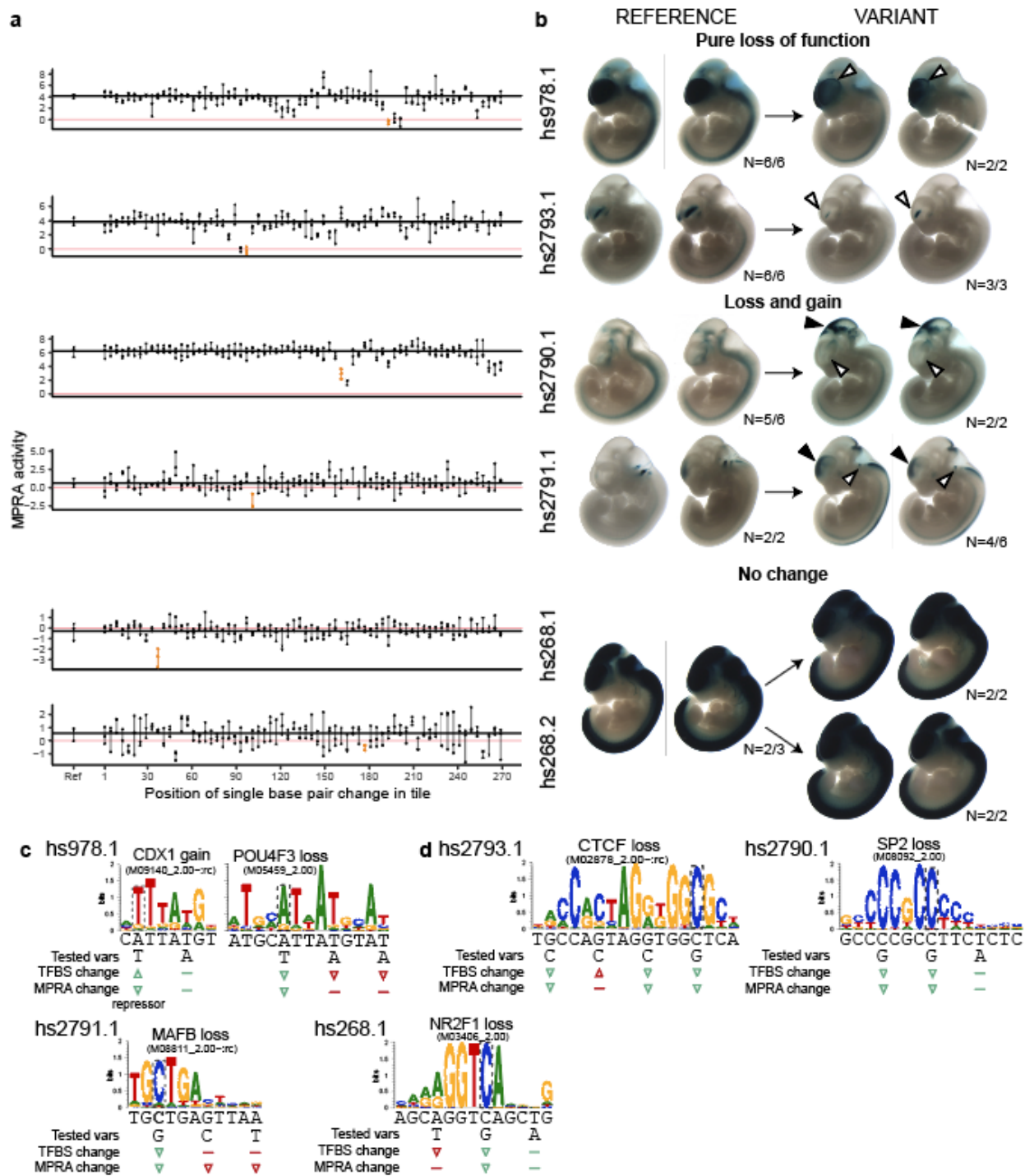
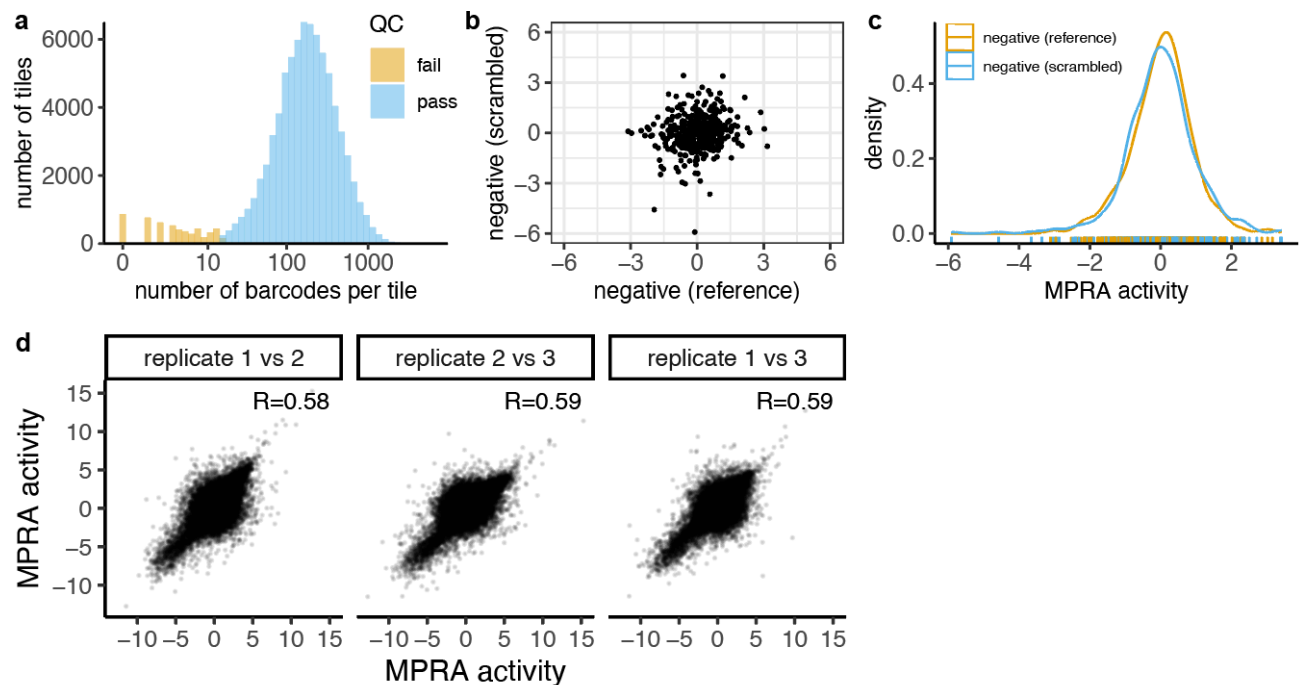


Figure 2.4: Synthetic MPRA variants lead to *in vivo* change of function in transgenic assay.

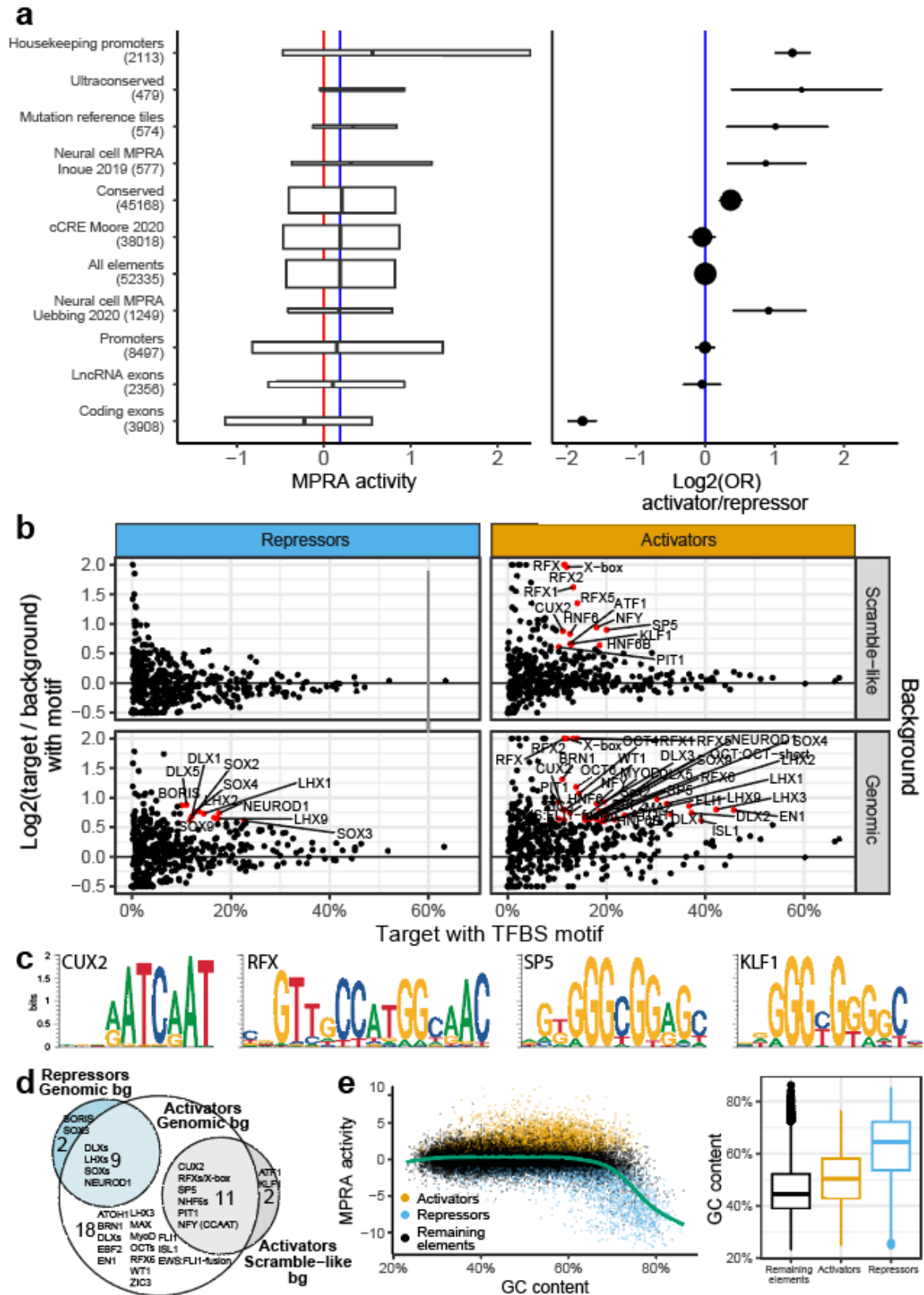
(a) Every fourth nucleotide of six MPRA tiles was mutagenized individually, for a total of 67 mutagenized constructs per reference tile. Dots connected by a vertical line = three biological replicates. Red horizontal line = zero, mean activity of scramble negative controls. Black horizontal line = mean activity of the reference construct. **(b)** Constructs (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
encompassing the MPRA tiles, with or without the variants indicated in orange in panel A, were tested for activity using transgenic assay in developing mouse embryos at embryonic day e11.5. All variants except hs268.1 and hs268.2 led to change of function in one or more neural tissues -brain, neural tube or cranial nerves. Shown are embryos that were genotyped as "tandem", i.e. positive for insertion at the safe harbor locus and presence of the plasmid backbone indicating multi-copy insertion with strong, reproducible pattern. White arrowhead indicates loss of function, black indicates gain of function. See Supplementary Figure 5 for all embryo images, which provide additional support of the changes observed when comparing tandem embryos. **(c)** Prediction of TFBS likely affected by the variants for hs978.1. Left: TFBS consistent with all MPRA variant effects (likely repressor CDX1), right: TFBS consistent with the MPRA effect of the variant tested in vivo, but with the effects of remaining two MPRA variants (POU4F3). TFBS and MPRA change symbols are colored green if matching and red if not. Arrowheads indicate an increase or decrease, flat line indicates no effect. These symbols assumes all TFBSs are activating, though we speculate CDX1 acts as a repressor. **(d)** Prediction of TFBS likely affected by the variants and partially or fully consistent with flanking variant effects.



Supplementary Figure 2.1: MPRA quality control.

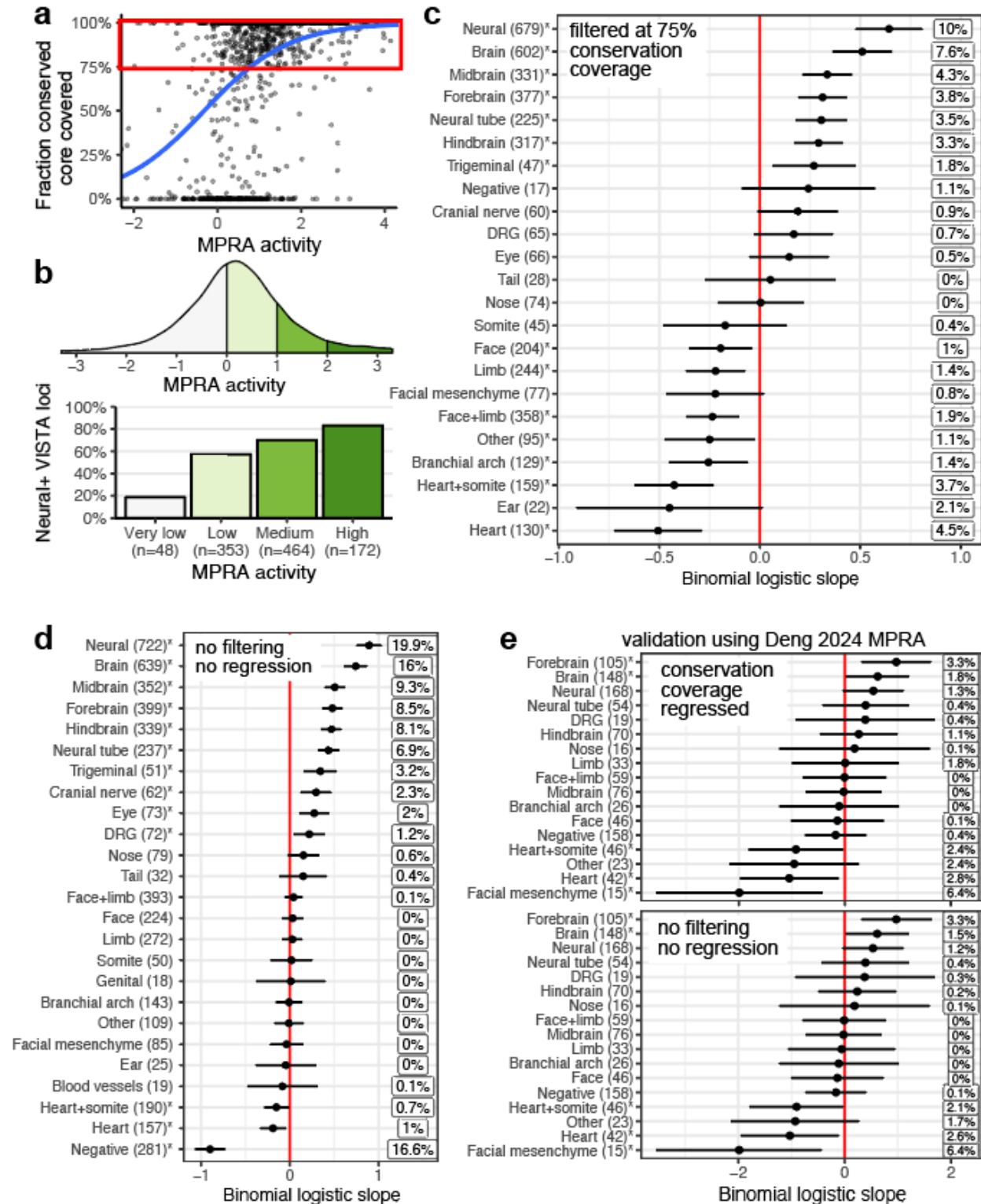
(a) Histogram of the number of barcodes per tile. (b) Scatterplot of MPRA activity comparing genomic reference negative elements and their dinucleotide scrambled equivalents. (c) Density plot of MPRA activity of genomic reference negative elements and their dinucleotide scrambled equivalents. Rug plot below indicates individual observations. (d) Correlation of MPRA activity between biological replicates. R -Pearson correlation.



Supplementary Figure 2.2: Neuronal WTC11 MPRA results validation.
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(a) Activity (left) and enrichment of significant activator/repressor tiles (right) across functional categories. Red line = activity of scrambled negative controls (zero, by definition). Blue line = activity (0.19) or log-odds-ratio of activators to repressor tiles (zero, by definition) of all reference elements. All categories in left panel have significantly different activity than scrambled negatives at FDR-adjusted p-value < 0.05 (Mann-Whitney U test). Bars in log-odds plot in the right panel are 95% confidence intervals - all categories with interval not overlapping 0 are significant at FDR-adjusted p-value < 0.05 (Fisher test). See Supplementary Table 1 for source data. Left panel is the same as **Figure 2a**. **(b)** TFBS enrichment in enhancer activator (N = 3,054) or repressor (N = 1,894) tiles compared to enhancer elements with scramble negative levels of activity (N = 15,503, "scramble-like") or genomic background elements (N = 50,000; "genomic"). Log₂-fold change was curbed at -0.5 and 2. Only TFBSs present in more than 10% target, with 50% increase in presence from background to target set (corresponding to $\log_2(1.5) = 0.58$ cutoff) and FDR < 1% are labeled. Top-right panel is the same as **Figure 2b**. **(c)** Examples of TFBSs enriched in enhancer activator tiles compared to enhancer elements with scramble negative levels of activity. **(d)** Overlap between TFBS enriched in different analyses from previous panel. TFs with similar names collapsed (e.g. "RFXs"). **(e)** Relationship between MPRA activity and GC content. Left: scatterplot. Green line is a smooth mean generated by a general additive model with REML parameter selection. Right: boxplot of GC content binned by tile category (activators N = 4,762, repressors N = 2,957, remaining elements N = 44,616). Hinges of boxplots span interquartile range (IQR), line in the middle is median, thickness (height) is proportional to number of overlapping tiles. Where used, whiskers extend from the hinge to the largest value no further than 1.5 * IQR from the hinge.

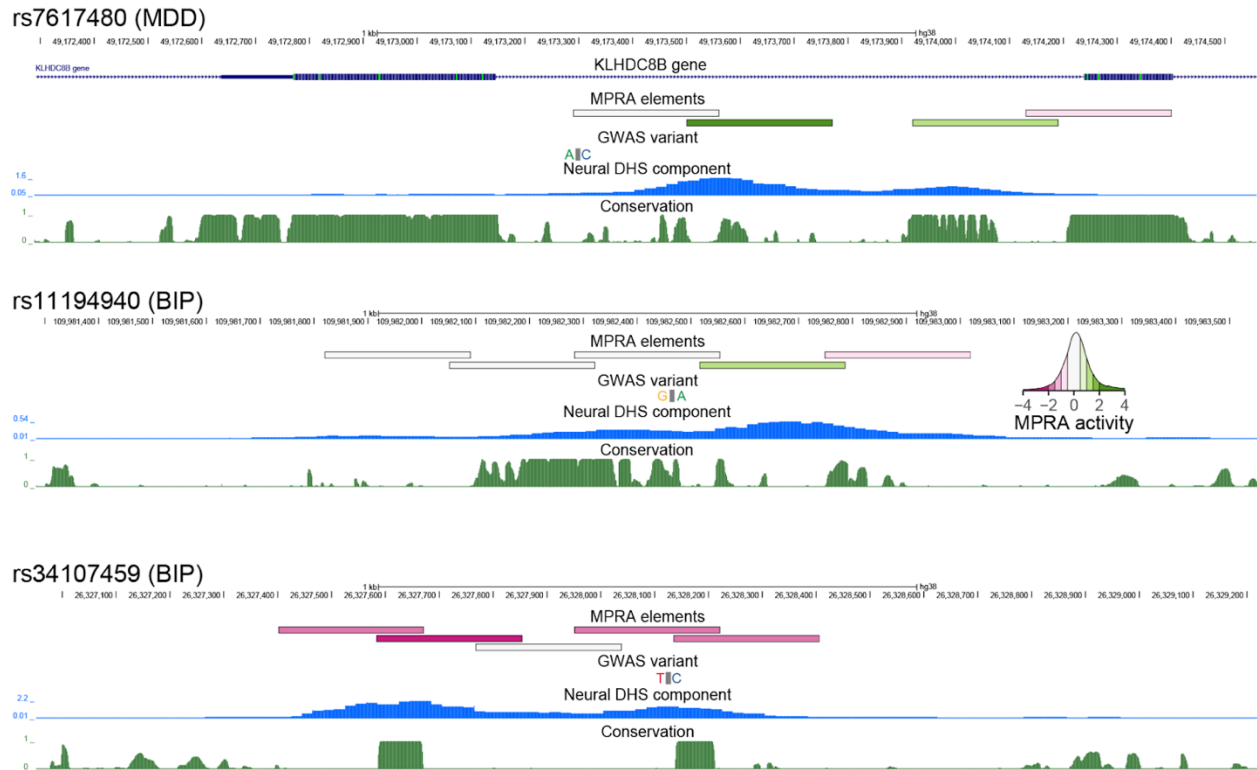


Supplementary Figure 2.3: Predicting transgenic assay activity using a MPRA-based, coverage-marginalized model.

(Figure caption continued on the next page.)

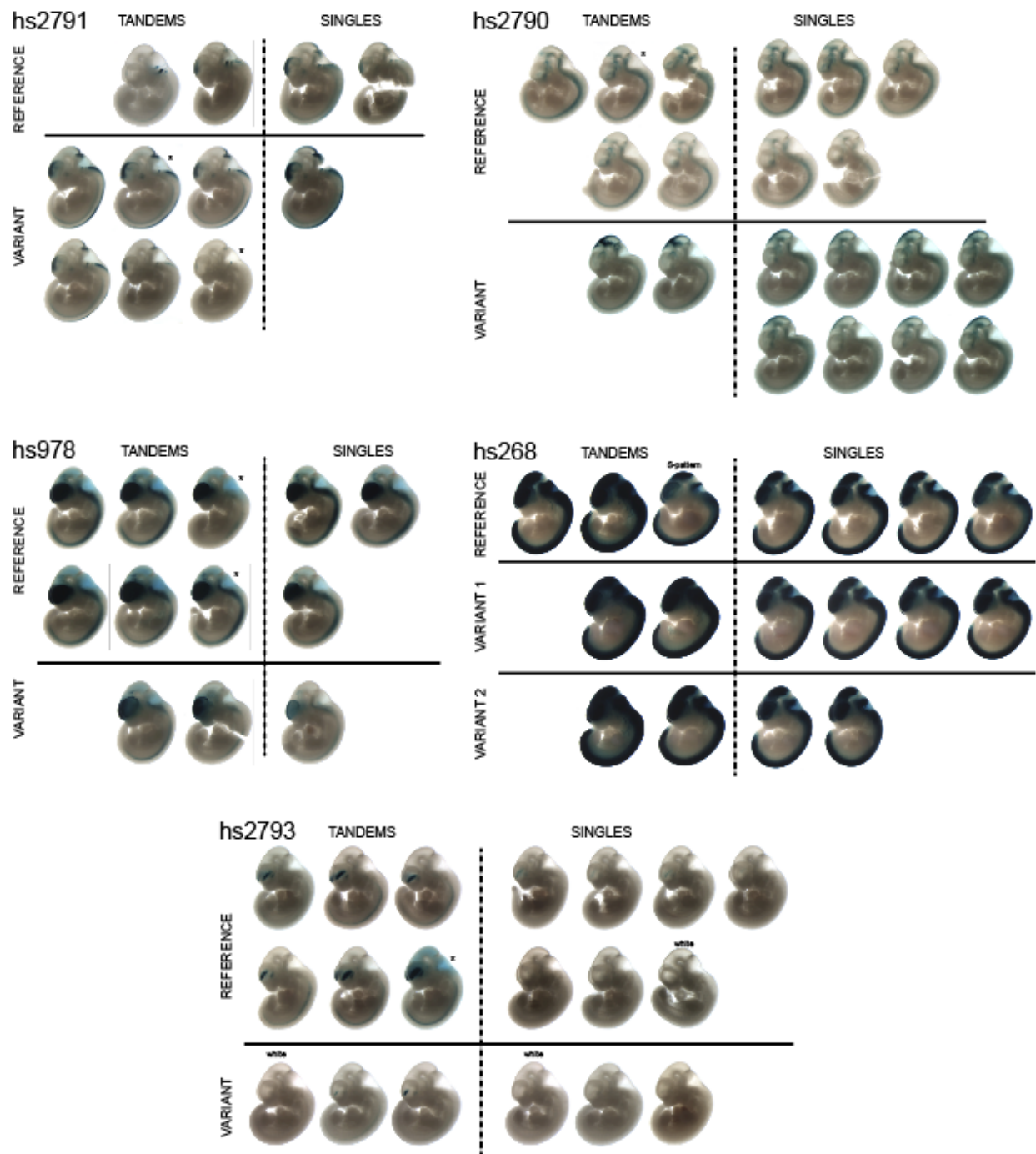
(Figure caption continued from the previous page.)

(a) Relationship between fraction of conserved core covered and MPRA activity. The blue line is the binomial-link GLM regression on this variable. Instead of including this variable as covariate in GLM, VISTA elements with coverage lower than 75% were removed prior to modeling. **(b)** Alternative visualization of relationship between MPRA activity and transgenic assay activity. Top: MPRA activity bins. Bottom: fraction of neural-positive VISTA loci by MPRA activity bin. Numbers below bars are counts of VISTA elements. Only "well-covered" elements included, as defined above (N = 1,037). **(c)** Results of the GLM predicting binomial transgenic assay activity of well-covered VISTA elements from MPRA activity. **(d)** Results of the GLM, without filtering or regression for conservation coverage. **(e)** GLM using Deng 2024 MPRA in primarily cortical cells. Asterisks indicate nominal p-value < 0.05. Boxed percentages to the right are Nagelkerke R^2 measures. Bars extend two standard errors of the mean in each direction. DRG = dorsal root ganglia. Cranial nerves category does not include the trigeminal nerve, as per VISTA Browser.



Supplementary Figure 2.4: Genomic tracks of the three nominally significant GWAS variants.

Conservation is PhastCons UCSC tracks for 30 mammals (27 primates) dataset. MPRA elements colored by MPRA activity, see inset. Neural DHS component from Meuleman 2020³². MDD = major depressive disorder. BIP = bipolar disorder.



Supplementary Figure 2.5: Results of transgenic mouse assay.

Tandems = embryos that were genotyped as positive for reporter integration at the safe harbor locus and presence of the plasmid backbone indicating higher transgene copy number with strong, reproducible pattern. Singles = embryos that were genotyped positive for reporter integration at the safe harbor locus and negative for plasmid backbone, indicating lower transgene copy number with weaker, but reproducible pattern. Asterisks - embryos with uncertain genotype. R-pattern - embryos with deviant (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
pattern indicative of random (R) genomic insertion. S-pattern - tandem embryos with expression pattern resembling that of single-genotyped embryos. Embryos without any staining are marked as "white".

Materials and Methods

MPRA design

We used following datasets for our library design - VISTA enhancers, MPRA tiles from Inoue 2019¹⁹ (activity > 1.1 at both 48h and 72h timepoints) and Uebbing 2021³⁶ ($q < 0.05$ for both replicates, following the publication) and single-cell or bulk ATAC or ATAC-seq peaks called by Ziffra 2020²⁶ (26,000 peaks designated enhancers by activity-by-contact), Domcke 2020²⁷ (33,000 cerebrum peaks with mean expression > 0.1), Preissl 2018²⁸ (top 15,000 peaks from each of eEX1, eIN1, eIN2 and RG1-4 clusters), Gorkin 2020²⁹ (top 15,000 from forebrain, midbrain, hindbrain and neural tube e11.5 samples), Inoue 2019¹⁹ (top 15,000 peaks from 72h timepoint) and Song 2019³⁰ (WTC11 neurons; top 15,000 peaks). These elements were either extended to 270 bp (if shorter) or tiled in intervals of 270 bp with a minimum 30 bp overlap. We also designed tiles directly upstream of first exons of coding genes in Gencode v34 with neural ATAC signal (one tile per promoter), facing in the direction of transcription and avoiding overlap with FANTOM5 CAGE TSS peaks⁵⁶. Tiles centered on representative DHS elements with 'Neural', 'Organ devel. / renal' and 'Primitive / embryonic' annotation were added, if overlapping previously chosen elements³². Genomic negative control tiles (N = 500) were selected from sections of negative VISTA elements that were not conserved, not active in previous neural MPRA and did not overlap any cCREs³² or any of the peaks in ATAC datasets mentioned above. Finally, we used a weighted combination of evolutionary conservation (UCSC phastConsElements30way⁵⁷), lack of overlap with coding exons (Gencode⁵⁸) and promoters regions (Gencode and CAGE⁵⁶), neural VISTA activity, presence of a peak in multiple ATAC datasets, activity in previous

neural MPRAs^{19,36} and overlap with LD blocks from psychiatric disorder GWAS to select 56,387 hg38 genomic reference tiles. We also used di-nucleotide scrambled 500 genomic negative control tiles to form scramble negative controls. All resources that were not originally available in hg38 (including mouse VISTA enhancers), were lifted over using Kent tools and relevant UCSC chain files⁵⁹. Design was conducted in R 4.3.2 with tidyverse 2.0.0 package^{60,61}.

We introduced mutations into 595 reference tiles, resulting in 123 tiles with multiple SNVs (derived from random mutagenesis of ultraconserved VISTA elements⁶²) and 24,942 with individual SNVs. To select GWAS SNPs for testing, we started with a set of 465 common lead SNPs from psychiatric disorder GWAS^{1,3-5} and extracted 15,133 SNPs in linkage disequilibrium (LD) with these using SNI_{PA} ($r^2 > 0.8$, 1000 genomes set, $v3$). We selected 186 for testing based on overlap with enhancers with high likelihood of activity (overlapping ATAC-seq peaks from multiple datasets, evolutionary conserved, active in transgenic assay or highly active in previous neuronal MPRAs). To investigate vulnerabilities of regulatory elements associated with GWAS signals, we conducted systematic mutagenesis of every fourth nucleotide in 85 tiles within 0.8 LD regions for additional 5,621 SNVs using a GC-preserving transversion scheme (G=C, A=T). Finally, we conducted similar systematic mutagenesis of 254 tiles with high likelihood of activity for an additional 17,272 GC-preserving transversion SNVs. Numbers of SNVs listed above are mutually exclusive, but some SNVs belonged to more than one category. For example, a total of 5,892 SNVs were in 0.8 LD regions, including synthetic, lead and LD SNPs. Note that about 10% of all designed elements were not successfully tested - see Results section for numbers after QC.

LentiMPRA cloning and infection

The lentiMPRA library was constructed as previously described¹⁵. A synthesized TWIST oligo pool with 300 bp long elements (270 bp insert + 2*15 bp PCR handles) was amplified by 12-cycle PCR using NEBNext High-Fidelity 2x PCR Master Mix (New England BioLabs, M0541L), the forward primer 5BC-AG-f01 and reverse primer 5BC-AG-r01 were used to add the minimal promoter, spacer and vector overhang sequence. The amplified fragment was purified using 1x of the HighPrep PCR Clean-up System (Magbio, AC-60500). The purified fragment underwent a second round of 12-cycle PCR using NEBNext High-Fidelity 2x PCR Master Mix (New England BioLabs, M0541L), the forward primer 5BC-AG-f02 and the reverse primer 5BC-AG-r02. This step added a 15 bp random barcode downstream from the minimal promoter. The amplified fragment was purified using Nucleospin Gel and PCR-Clean-Up (Macherey-Nagel, 740609.50) and 1.2x HighPrep PCR Clean-up System (Magbio, AC-60500). The oligo library was cloned into the double digested *Agel/SbfI* pLS-*Scel* vector (Addgene, 137725) using the NEBuilder HiFi DNA Assembly Master Mix (New England BioLabs, E2621L). The plasmid lentiMPRA library was electroporated into MegaX DH10B T1R Electrocomp Cells (Invitrogen, C640003) using the Gemini X2 (2.0 kV, 200 Ω , 25 μ F). The electroporated cells were then plated on eleven 15 cm 100 mg/mL ampicillin LB agar plates (Teknova, L5004) and grown overnight at 37 °C. Approximately 8 million colonies were pooled and midi-prepped (Qiagen, 12145) to obtain on average 100 barcodes per oligo. To associate barcodes with each oligo in the library, the Illumina flow cell adapters were added through a 15-cycle PCR using NEBNext High-Fidelity 2x PCR Master Mix (New England BioLabs, M0541L), the forward primer pLSmP-ass-i741 and

reverse primer pLSmP-ass-gfp. The amplified fragment was purified using Nucleospin Gel and PCR-Clean-Up (Macherey-Nagel, 740609.50) and 1.8x HighPrep PCR Clean-up System (Magbio, AC-60500). The amplified fragments were sequenced on a NovoSeq 500 using a NextSeq 150PE kit with custom primers (R1: pLSmP-ass-seq-R1, R2: pLSmP-ass-seq-ind1, R3: pLSmP-ass-seq-R2).

Lentivirus production was conducted on twenty-nine 10 cm dishes of LentiX 293T cell line (TakaraBio, 632180) with Lenti-Pac HIV expression packaging kit (GeneCopoeia, LT002) following the manufacturer's protocol. Lentivirus was filtered through a .45 µm PES filter system (Thermo Fisher Scientific, 165-0045) and concentrated with Lenti-X Concentrator (TakaraBio, 631232). Titration of the lentiMPRA library was conducted on differentiated human excitatory neurons. Cells were seeded at 4.5×10^4 cells per well in a 12-well plate on day 0 and incubated for 7 days. Serial volumes of the lentivirus (0, 1, 2, 4, 8, 16, 32, 64, 128 µL) were added along with 6 µL ViroMag R/L (OZ Biosciences, RL41000) per well. After lentivirus addition cells were incubated for 30 minutes on the magnet at 37 °C. The magnet was removed and cells were incubated at 37 °C for 7 days, the media was replaced after 24 hours of lentivirus addition. The cells were washed with DPBS (Sigma-Aldrich, D8537) and DNA was extracted with the AllPrep DNA/RNA Mini Kit (Qiagen, 80204) following the manufacturer's protocol for DNA extraction. The multiplicity of infection (MOI) was determined as the relative amount of viral DNA over that of genomic DNA by qPCR using SsoFast EvaGreen Supermix (Bio-Rad, 1725205). The lentivirus infection, DNA/RNA extraction and DNA/RNA barcode sequencing were conducted as previously described¹⁵. Each replicate required approximately 25 million cells. Therefore, cells

were seeded at day 0 of differentiation in four 10 cm plates with 5 million cells each. On day 7, the cells were infected with the lentivirus library and ViroMag R/L (OZ Biosciences, RL41000) following the manufacturer's protocol. All three replicates were infected with the same lentivirus batch at an MOI of 80. Media was replaced 24 hours after lentivirus addition and the cells were incubated for 7 days. DNA and RNA were extracted from the three replicates using the AllPrep DNA/RNA Mini Kit (Qiagen, 80204) following the manufacturer's protocol. The RNA was treated with the TURBO DNA-free Kit (Life Technologies, AM1907) following the manufacturer's protocol for rigorous DNase treatment. Reverse transcription was conducted with SuperScript II Reverse Transcriptase (Life Technologies, 18064-071) using a barcode-specific primer (P7-pLSmP-ass16UMI-gfp) which contains a 16 bp UMI. After DNase treatment and reverse transcription the resulting cDNA and extracted DNA underwent the same steps to prepare the library for sequencing. To add a sample index and UMI, DNA and cDNA from the three replicates were kept separate and underwent a 3-cycle PCR using NEBNext Ultra II Q5 Master Mix (New England Biolabs, M0544L), forward primer P7-pLSmp-ass16UMI-gfp and reverse primer P5-pLSmP-5bc-i#. Another round of PCR was conducted to prepare the library for sequencing using NEBNext Ultra II Q5 Master Mix (New England Biolabs, M0544L), forward primer P5 and reverse primer P7. The fragments were purified using 1.2x of the HighPrep PCR Clean-up System (Magbio, AC-60500). The final libraries were sequenced on four runs of Illumina NextSeq high-output using the custom primers (R1: pLSmP-ass-seq-ind1, R2: pLSmP-UMI-seq, R3: pLSmP-bc-seq, R4: pLSmP-5bc-seq-R2).

Cell culture and neuronal differentiation

Differentiated human excitatory neurons were derived from hiPSCs in the WTC11 background where a doxycycline-inducible neurogenin 2 transgene was integrated into the AAVS1 locus³⁴. In the undifferentiated stage, cells were maintained in mTeSR 1 (STEMCELL Technologies, 85850) and the medium was changed daily. Once confluent, cells were washed with 1x DPBS (Sigma-Aldrich, D853), dissociated with accutase (STEMCELL Technologies, 07920) and plated at a 1:6 ratio in matrigel (Corning, 354277) coated plates. Media was supplemented with ROCK inhibitor Y-27632 (STEM CELL, 72304) at 10 μ M on the day of passage. To initiate differentiation, cells were washed with 1x DPBS, dissociated with accutase and plated in matrigel-coated plates. For three days cells were cultured in KnockOut DMEM/F-12 (Life Technologies, 12660-012) medium supplemented with 2 μ g/mL doxycycline (Sigma-Aldrich, D9891), 1X N-2 Supplement (Life Technologies, 17502-048), 1X NEAA (Life Technologies, 11140-050), 10 ng/mL BDNF (PeproTech, 450-02), 10 ng/mL NT-3 (PeproTech, 450-03) and 1 μ g/mL lamininin (Life Technologies, 23017-015). The pre-differentiation medium was changed daily for three days and on the first day medium was supplemented with 10 μ M ROCK inhibitor Y-27632. To induce neuronal maturation, cells were lifted and plated in Poly-L-Ornithine (Sigma-Aldrich, P3655) coated plates. Cells were cultured in maturation media containing Neurobasal A (Life Technologies, 12349-015) and DMEM/F12, HEPES (Life Technologies, 11330-032) with 2 μ g/mL doxycycline supplemented with 1X N-2 Supplement, 0.5X B-27 Supplement, 1X NEAA, 0.5X GlutaMax (Life Technologies, 35050-061), 10 ng/mL BDNF, 10 ng/mL NT-3 and 1

µg/mL lamininin. A half-media change was conducted on day 7 and day 14 of differentiation using the maturation medium minus doxycycline.

LentiMPRA analysis

Processing of barcode association and final MPRA libraries was done using a standardized MPRAflow pipeline^{15,35,38}, without a MAPQ filter to avoid artificial dropout due to multi-mapping of elements with single base pair mutations. All subsequent analyses were conducted in R 4.3.2 with tidyverse 2.0.0 package⁶¹. Visualizations were done using ggrastr 1.0.2 (<https://github.com/VPetukhov/ggrastr>), ggplot2 3.5.0⁶³ and ggrepel 0.9.5⁶⁴. General linear models were constructed using rms 6.8-0 (<https://hbiostat.org/r/rms/>). Motifs affected by variants tested in the transgenic assay were detected using motifbreakR 2.16.0⁴⁹ with filterp=T, threshold=1e-4 and pwmList from Viestra 2020⁶⁵. Only tiles with at least 15 barcodes detected in each of the three replicates were retained and mutation tiles without a reference passing these criteria were discarded as well. As per MPRAflow pipeline, these barcodes include ones detected in DNA or RNA. In other words, barcodes detected using only one modality were not discarded. MPRA activity was expressed as a z-score of $\log_2(\text{RNA counts}/\text{DNA counts})$ relative to scramble negative controls.

Correlation of MPRA activity and epigenomic signal

Epigenomic signal in the form of bigWig files was retrieved from ENCODE³² and 12 other sources (**Supplementary Table 2.5**) and integrated over tile intervals using *bedtools bigWigAverageOverBed* command⁶⁶. For each sample, signal was sorted and ranked with random tie-breaking. For a range of rank cutoffs starting with 1000, the tiles were split into those above and below the cutoff and median MPRA activity was

computed for both groups. The median activity of bottom signal group (e.g. from rank 1001 to lowest rank) was then subtracted from median activity of the top signal group (e.g. ranks 1-1000). For enhancer analysis, 8495 tiles overlapping promoters defined as 2 kb centered on the 5' end of exon 1 of protein-coding genes in Gencode V34⁵⁸, were removed before computing the ranks and median activity difference.

TFBS enrichment analysis

All analysis was done using HOMER 4.11⁴¹ using activator or repressor tiles as target (as defined in the main text) and either HOMER-selected, GC-matched background genomic elements of the same size, or library elements with scramble negative levels of activity ($-0.4 < \text{activity} < 0.4$). Only tiles not overlapping promoters, as defined in the previous section, were used. Default set of 239 unique TF motifs was used. Command of the form *findMotifsGenome.pl target.bed hg38 target_folder -bg background.bed -size 270 -nomotif* was run for each analysis, except *-bg* term was dropped for HOMER-selected background.

Alignment and preprocessing of functional genomic data for ABC score pipeline

Gestational week 18 (GW18) bulk ATAC-seq and H3K27ac ChIP-seq data from human fetal prefrontal cortex⁶⁷ were aligned to hg19 using the standard Encode Consortium ATAC-seq and ChIP-seq pipelines respectively with default settings and pseudo replicate generation turned off (<https://github.com/ENCODE-DCC>). Trimmed, sorted, duplicate and chrM removed ATAC-seq and sorted, duplicate removed ChIP-seq bam files produced by the Encode pipeline were provided as input for calculating ABC scores. ATAC-seq and H3K27ac CUT&RUN data from 7-8 week old NGN2-iPSC inducible excitatory neurons was obtained from Song 2019³⁰. ATAC-seq and CUT&RUN

reads were trimmed to 50 bp using TrimGalore⁶⁸ with the command `--hardtrim 5 50` before alignment. ATAC-seq reads were aligned to hg19 using the standard Encode Consortium ATAC-seq and ChIP-seq pipelines respectively with default settings and pseudo replicate generation turned off. Trimmed, sorted, duplicate and chrM removed ATAC-seq bam files from multiple biological replicates were combined into a single bam file using `samtools merge v1.10`⁶⁹. Trimmed CUT&RUN reads were aligned to hg19 using Bowtie2 v2.3.5.1⁷⁰ with the following settings `--local --very-sensitive-local --no-mixed --no-discordant -l 10 -X 700` and output sam files were converted to bam format using `samtools view`^{69,70}. Duplicated reads were removed from the CUT&RUN bam file using Picard MarkDuplicates v2.26.0⁷¹ with the `--REMOVE_DUPLICATES =true` and `--ASSUME_SORTED=true` options (<http://broadinstitute.github.io/picard/>). The final ATAC-seq and CUT&RUN bam files were provided as input for calculating ABC scores.

Preprocessing of HiC and pHiC data for ABC score pipeline

HiC contacts with 10 kb resolution from human GW17-18 fronto-parietal cortex was obtained in an hdf5 format separated by chromosome⁷²(**Supplementary Table 2.6**). Hdf5 files were filtered for contacts with a score > 0 and converted into a bedpe format. Promoter capture HiC (pHiC) contacts from 7-8 week old NGN2-iPSC inducible excitatory neurons were obtained in an ibed format from GSE113483³⁰. The ibed file was converted to bedpe format and separated by chromosome. Bedpe files from GW17-18 cortex and iPSC derived excitatory neurons were provided as input for calculating ABC scores.

Identification of candidate enhancer-gene pairs with ABC Score

The Activity-by-Contact (ABC) model identifies enhancer-gene relationships based on chromatin state and conformation⁴⁸. Previously identified open chromatin regions from GW18 human prefrontal cortex⁶⁷ and corresponding ATAC-seq and H3K27ac ChIP-seq bam files were provided as input for the ABC score pipeline `MakeCandidateRegions.py` script with the flags `--peakExtendFromSummit 250 --nStrongestPeaks 150000`. Candidate enhancer regions identified were then provided to the `run.neighborhoods.py` script in addition to hg19 promoter merged transcript bounds. Finally, `predict.py` was used to identify final candidate enhancers using HiC data from human GW17-18 fronto-parietal cortex with the flags `--hic_type bedpe --hic_resolution 10000 --scale_hic_using_powerlaw --threshold .02 --make_all_putative72`. Candidate enhancer-gene pairs were also identified for 7-8 week old NGN2-iPSC inducible excitatory neurons using respective open chromatin regions³⁰, ATAC-seq and H3K27ac ChIP-seq data. All other settings for the ABC score pipeline remained constant.

Mouse enhancer transgenic assay

Transgenic E11.5 mouse embryos were generated as described previously¹⁶. Briefly, super-ovulating female FVB mice were mated with FVB males and fertilized embryos were collected from the oviducts. Regulatory elements sequences were synthesized by Twist Biosciences. Inserts generated in this way were cloned into the donor plasmid containing minimal *Shh* promoter, *lacZ* reporter gene and H11 locus homology arms (Addgene, 139098) using NEBuilder HiFi DNA Assembly Mix (NEB, E2621). The sequence identity of donor plasmids was verified using long-read sequencing (Primordium). Plasmids are available upon request. A mixture of Cas9

protein (Alt-R SpCas9 Nuclease V3, IDT, Cat#1081058, final concentration 20 ng/μL), hybridized sgRNA against H11 locus (Alt-R CRISPR-Cas9 tracrRNA, IDT, cat#1072532 and Alt-R CRISPR-Cas9 locus targeting crRNA, gctgatggaacaggtaacaa, total final concentration 50 ng/μL) and donor plasmid (12.5 ng/μL) was injected into the pronucleus of donor FVB embryos. The efficiency of targeting and the gRNA selection process is described in detail in Osterwalder 2022¹⁶. Embryos were cultured in M16 with amino acids at 37°C, 5% CO₂ for 2 hours and implanted into pseudopregnant CD-1 mice. Embryos were collected at E11.5 for lacZ staining as described previously¹⁶. Briefly, embryos were dissected from the uterine horns, washed in cold PBS, fixed in 4% PFA for 30 min and washed three times in embryo wash buffer (2 mM MgCl₂, 0.02% NP-40 and 0.01% deoxycholate in PBS at pH 7.3). They were subsequently stained overnight at room temperature in X-gal stain (4 mM potassium ferricyanide, 4 mM potassium ferrocyanide, 1 mg/mL X-gal and 20 mM Tris pH 7.5 in embryo wash buffer). PCR using genomic DNA extracted from embryonic sacs digested with DirectPCR Lysis Reagent (Viagen, 301-C) containing Proteinase K (final concentration 6 U/mL) was used to confirm integration at the H11 locus and test for presence of tandem insertions¹⁶. Only embryos with donor plasmid insertion at H11 were used. The stained transgenic embryos were washed three times in PBS and imaged from both sides using a Leica MZ16 microscope and Leica DFC420 digital camera.

References

1. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
2. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
3. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu & Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469–1482.e11 (2019).
4. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
5. Mullins, N. *et al.* GWAS of Suicide Attempt in Psychiatric Disorders and Association With Major Depression Polygenic Risk Scores. *Am. J. Psychiatry* **176**, 651–660 (2019).
6. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–10 (2015).
7. Goes, F. S. *et al.* De novo variation in bipolar disorder. *Mol. Psychiatry* **26**, 4127–4136 (2021).
8. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
9. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in

- vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
10. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
 11. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
 12. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
 13. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
 14. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
 15. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
 16. Osterwalder, M. *et al.* Characterization of Mammalian In Vivo Enhancers Using Mouse Transgenesis and CRISPR Genome Editing. *Methods Mol. Biol.* **2403**, 147–186 (2022).
 17. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262–1271.e15 (2020).
 18. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**,

- D88–D92 (2006).
19. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* **25**, 713–727.e10 (2019).
 20. Caputo, D. *et al.* Characterization of enhancer activity in early human neurodevelopment using Massively parallel reporter assay (MPRA) and forebrain organoids. *bioRxiv* (2023) doi:10.1101/2023.08.14.553170.
 21. Whalen, S. *et al.* Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron* **111**, 857–873.e8 (2023).
 22. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
 23. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
 24. Zeng, B. *et al.* Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases. *bioRxiv* (2023) doi:10.1101/2023.03.02.530826.
 25. Kreimer, A. *et al.* Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* **13**, 1504 (2022).
 26. Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
 27. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, (2020).
 28. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.*

- 21**, 432–439 (2018).
29. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
 30. Song, M. *et al.* Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019).
 31. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
 32. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
 33. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
 34. Wang, C. *et al.* Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening. *Stem Cell Reports* **9**, 1221–1233 (2017).
 35. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
 36. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
 37. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
 38. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of

- massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
39. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
 40. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
 41. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
 42. Sugiaman-Trapman, D. *et al.* Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* **19**, 181 (2018).
 43. Leung, R. F. *et al.* Genetic Regulation of Vertebrate Forebrain Development by Homeobox Genes. *Front. Neurosci.* **16**, 843794 (2022).
 44. Landry, C. *et al.* HNF-6 is expressed in endoderm derivatives and nervous system of the mouse embryo and participates to the cross-regulatory network of liver-enriched transcription factors. *Dev. Biol.* **192**, 247–257 (1997).
 45. Stevanovic, M. *et al.* SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. *Front. Mol. Neurosci.* **14**, 654031 (2021).
 46. Yang, X. *et al.* Functional characterization of gene regulatory elements and neuropsychiatric disease-associated risk loci in iPSCs and iPSC-derived neurons. *bioRxiv* 2023.08.30.555359 (2023) doi:10.1101/2023.08.30.555359.
 47. Deng, C. *et al.* Massively parallel characterization of regulatory elements in the

- developing human cortex. *Science* **384**, eadh0559 (2024).
48. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
 49. Coetzee, S. G. & Hazelett, D. J. motifbreakR: A Package for Predicting the Disruptiveness of Single Nucleotide Polymorphisms on Transcription Factor Binding Sites. *Bioconductor version: Release (3.12)*.
 50. Huang, D. *et al.* The role of Cdx2 as a lineage specific transcriptional repressor for pluripotent network during the first developmental cell lineage segregation. *Sci. Rep.* **7**, 17156 (2017).
 51. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *Elife* **10**, (2021).
 52. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11952–11957 (2013).
 53. Deng, C. *et al.* Massively parallel characterization of psychiatric disorder-associated and cell-type-specific regulatory elements in the developing human cortex. *bioRxiv* (2023) doi:10.1101/2023.02.15.528663.
 54. Hollingsworth, E. W. *et al.* Rapid and Quantitative Functional Interrogation of Human Enhancer Variant Activity in Live Mice. *bioRxiv* (2023) doi:10.1101/2023.12.10.570890.
 55. Levo, M. *et al.* Transcriptional coupling of distant regulatory genes in living embryos. *Nature* **605**, 754–760 (2022).

56. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
57. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
58. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
59. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
60. Ripley, B. D. The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network* **1**, 23–25 (2001).
61. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
62. Snetkova, V. *et al.* Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.* **53**, 521–528 (2021).
63. Wickham, H. Getting Started with ggplot2. in *ggplot2: Elegant Graphics for Data Analysis* (ed. Wickham, H.) 11–31 (Springer International Publishing, Cham, 2016).
64. Slowikowski, K. ggrepel: Automatically position non-overlapping text labels with ‘ggplot2’. *R package version 0.9.1*, (2021).
65. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
66. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–34 (2014).
67. Markenscoff-Papadimitriou, E. *et al.* A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell* **182**, 754–769.e18 (2020).

68. Krueger, F. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *Babraham Institute* (2015).
69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
71. Picard toolkit. *Broad Institute, GitHub repository* (2019).
72. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

CHAPTER 3:

FUNCTIONAL CHARACTERIZATION OF *OXTR*-ASSOCIATED ENHANCERS.

Abstract

The oxytocin receptor (*OXTR*) has a vital role in regulating human behavior, controlling lactation, parturition, pair bonding, maternal behavior, anxiety, and sociability. However, its regulatory elements and how variation in these sequences lead to behavioral changes remain largely unknown. Here, we identified seven *OXTR* candidate *cis*-regulatory elements (cCREs) from mouse and human hypothalamus single-cell RNA/ATAC-seq data and characterized them in cells and mice. Luciferase assays in hypothalamus cell lines identified three of the seven to be functional enhancers. Mouse enhancer assays for the most robust enhancer, *OXTR* candidate enhancer 7 (OCE7), found it to be active in the mouse olfactory bulb at postnatal day 28 and day 56. In summary, using genomic data coupled with cell and mouse enhancer assays, we characterized the *OXTR* regulatory landscape and identified a novel olfactory bulb *OXTR*-associated enhancer.

Introduction

Oxytocin (OXT) and the oxytocin receptor (*OXTR*) play a vital role in lactation, parturition, pair bonding, maternal behavior, anxiety, autism spectrum disorder (ASD), and sociability¹. *OXTR* is the main target for oxytocin in the central nervous system (CNS) and the periphery, mediating the effects of oxytocin in regulating behavior. OXT is a neuropeptide synthesized in the magnocellular and parvocellular neurons of the paraventricular and supraoptic nuclei of the hypothalamus¹. OXT is stored in secretory vesicles and released from axonal terminals to the posterior lobe of the pituitary for

peripheral circulation². OXT can exert its function through OXTR in two ways: 1) by locally binding to the *OXTR* receptor and 2) through diffusion to the extracellular space where it can reach distant brain areas and the periphery where *OXTR* is expressed². *OXTR* is a class of G protein-coupled receptors (GPCRs) that when bound by its ligand, oxytocin can initiate several signaling cascades¹. In contrast to oxytocin, *OXTR* is expressed throughout the brain and periphery, where it controls the final rate-limiting step of oxytocin signal transmission, thereby mediating the pivotal role of oxytocin in regulating maternal behavior, parturition, lactation, and social behaviors.

Over the last two decades, studies have looked at the association of *OXTR* with psychiatric disorders and behavior. For example, one study described a 0.7 Mb rare deletion at 3p25.3 encompassing the *OXTR* gene alongside four other genes (*LMCD1*, *CAV3*, *RAD18*, and *C3orf32*) in a patient with ASD³. In addition, several studies have analyzed the methylation patterns of *OXTR* finding associations with methylation patterns and anxiety, ASD, and depression⁴⁻⁷. Additionally, a study in 198 simplex ASD families demonstrated an association of rs237887 and face recognition deficits in patients with ASD⁸. There have also been large cohort studies focusing on the association of SNPs around *OXTR* to human prosocial behaviors. Most notably, a single nucleotide polymorphism (SNP) in intron 3 of *OXTR* (rs53576) has been shown to have significant association with empathy in humans^{9,10}. Overall, non-coding and coding variants in *OXTR* have been associated with ASD, developmental prosopagnosia, and empathy^{8,9,11-13}.

Rodent studies have also increased our understanding of the role of *Oxtr* in regulating social behaviors. Genetically modified mouse models of oxytocin and the

oxytocin receptor demonstrated pervasive social deficits. *Oxt* null mice (*Oxt*^{-/-}) were found to have pervasive social recognition deficits¹⁴. Similarly, *Oxtr* homozygous knockout mice (*Oxtr*^{-/-}) demonstrated impairments in social discrimination, changes in maternal behavior, and increased aggression¹⁵. Heterozygous knockout *Oxtr* mice exhibit impaired sociability and preference for social novelty, but unlike the homozygous genotype, their cognitive flexibility and aggression were normal¹⁶. As heterozygous *Oxtr* mice showed a behavioral phenotype, this indicates that *Oxtr* is dosage sensitive, suggesting that mutations in *cis*-regulatory elements (cCREs) controlling *Oxtr* expression might lead to an *Oxtr* dose-dependent behavioral phenotype.

Not much is known about the regulatory landscape of *OXTR*. A better understanding of the regulatory network of *OXTR* will provide insights into where *OXTR* is expressed in the human brain and how its alteration can lead to behavioral changes or be utilized as a therapeutic target¹⁷. In this study, we took advantage of genomic datasets to identify cCREs of *OXTR* and characterize their function *in vitro* and *in vivo*. Testing seven cCREs in mouse hypothalamus cells found three sequences to show significant enhancer activity including, OCE7, that had the most robust enhancer activity. Previous studies have pointed at the possibility of a *cis*-regulatory element present in OCE7^{18,19}, but no study has functionally validated this region. Carrying out a transgenic mouse enhancer assay for this sequence, we found it to be an active enhancer in the mouse olfactory bulb. Combined, our work annotated *OXTR* associated cCREs and found three functional enhancer sequences *in vitro* and one functional olfactory bulb enhancer.

Results

Annotation of OXTR cCREs.

We used a combination of comparative and functional genomic datasets to identify *OXTR* candidate enhancer sequences. First, we selected *OXTR* candidate enhancer sequences that are within the *OXTR* topologically associated domain (TAD) boundary. We annotated the TAD boundary of *OXTR* using mouse and human publicly available Hi-C data from neural cell types accessible via the 3D Genome Browser²⁰. Moreover, we made use of a recent study that described the mouse and human *OXTR* TAD boundaries²¹. Secondly, we assessed sequence conservation using the UCSC Genome Browser “Vertebrate Multiz Alignment & Conservation (100 Species)” track²². Third, we selected sequences that demonstrate active enhancer marks, as defined by the presence of EP300 and histone 3 lysine 27 acetylation (H3K27ac) chromatin immunoprecipitation followed by sequencing (ChIP-seq) peaks, from ENCODE and psychENCODE datasets^{23,24}. Fourth, we used mouse and human hypothalamus scRNA-seq/scATAC-seq generated from our lab²⁵. Through this methodology, we identified seven *OXTR* candidate enhancers (OCEs) that were selected for subsequent luciferase assays (**Figure 3.1A, Supplementary Table 3.1**).

Luciferase assays of OXTR cCREs identify three functional enhancers.

We cloned all seven OCEs into an enhancer assay vector (pGL4.23; Promega) in front of a minimal promoter and the luciferase reporter gene (**Supplementary Table 3.1**). The vectors were individually transfected into mouse hypothalamus cells, mHypoA-POMC, along with a Renilla luciferase vector (pGL4.74; Promega) to correct for transfection efficiency. Luciferase activity was measured after 48 hours of

transfection. From the seven *OXTR* candidate enhancer sequences, we identified three OCEs to have significant enhancer activity: OCE1, OCE5 and OCE7 (**Figure 3.1b**). Two of these OCEs, OCE1 and OCE5, were derived from the mouse and human hypothalamus scRNA-seq/scATAC-seq. These sequences also overlapped neural DNase-I hypersensitive sites. Moreover, OCE1 also overlapped two of the ENCODE registry of candidate cCREs.

The candidate enhancer sequence with the highest luciferase activity was OCE7 (**Figure 3.1C**). OCE7 is located in intron 3 of the *OXTR* gene, and this intron has been previously suggested to contain cCREs that regulate *OXTR* expression^{18,19,21}. SNPs near this region have been associated with empathy, ASD, and developmental prosopagnosia. (i.e., rs53576, rs237887, and rs2254298)^{8,9,11}. Moreover, using the Single nucleotide polymorphisms annotator (SNIIPA) we determined that rs2254298, which is associated with developmental prosopagnosia, is in linkage disequilibrium with rs2268491 ($R^2=0.98$), located in OCE7 and which has also been associated with ASD (**Supplementary Table 3.2**)^{10,26}. To test the effect of rs2268491 on OCE7 enhancer activity, we generated an OCE7 enhancer assay vector containing the alternate allele (T) at position chr3:8758712 (hg38) via site-directed mutagenesis. We next carried out a similar luciferase assay in mHypoA-POMC cells using both reference and alternate allele vectors. We observed that the alternate allele (T) significantly increased enhancer activity compared to the reference allele (C) (**Figure 3.1D**). In summary, our luciferase assays identified three functional *OXTR* cCREs in mouse hypothalamus cells, with OCE7 driving the most robust enhancer activity. In addition, we found that the alternate allele of rs2268491, that is associated with ASD, leads to an increase in OCE7

enhancer activity.

OCE7 is an active enhancer in the mouse olfactory bulb.

We next used mouse enhancer assays to characterize the spatiotemporal activity of OCE7, as it was the most robust enhancer in our luciferase assays. We cloned OCE7 into a mouse enhancer assay vector containing the heat shock protein 68 (*Hsp68*) minimal promoter followed by the mCherry reporter gene²⁷. We generated three different founder lines and used qRT-PCR to determine transgene copy number in these lines. The three independent lines were found to contain varying copy numbers of the transgene, ranging from 2 transgene copies in line 1 to 52 transgene copies in line 2 (**Figure 3.2A**). We used these three independent lines for all subsequent assays. We assayed enhancer activity via immunofluorescence in postnatal mouse brains at different time points (P28 and P56) in both the hypothalamus and olfactory bulb, which are known to express *Oxtr*²⁸⁻³⁰. The time points selected were guided by key developmental time points where *Oxtr* has been studied previously in rodents²⁸, selecting adult time points (P28 and P56). We observed strong enhancer activity at P28 and P56 in the olfactory bulb in all three mouse lines using qRT-PCR, fluorescence imaging, and immunohistochemistry using antibodies for mCherry (**Figure 3.2B-C**, **Supplementary Figure 3.1A-B**). In addition, careful evaluation of the hypothalamus did not find any detectable enhancer activity at P28 or P56 (**Supplementary Figure 3.2A-D**). In summary, our results suggest that OCE7 is an active enhancer in the olfactory bulb.

Discussion

Using functional genomics coupled with mouse enhancer assays, we characterize a novel olfactory bulb enhancer in the *OXTR* locus. Although previous studies have pointed to the possibility of an enhancer element in this region^{18,19,21}, no study has thoroughly characterized its enhancer activity. Moreover, we identified seven *OXTR* candidate enhancer regions in the *OXTR* locus derived from functional genomics datasets. We functionally tested these sequences *in vitro* using luciferase assays in mouse hypothalamus cells identifying three sequences to have significant enhancer activity (OCE1, OCE5 and OCE7). We determined the alternate allele of rs2268491 to significantly increase OCE7 enhancer activity. To further characterize the spatiotemporal expression of OCE7, we generated stable transgenic mouse enhancer lines finding it to function as an active olfactory bulb enhancer.

Out of the seven OCEs that we characterized, three had significant enhancer activity in mouse hypothalamus cell lines. Due to time and cost limitations, we only characterized the most robust of these enhancers, OCE7 in mice. Surprisingly, despite obtaining strong enhancer activity in the mouse hypothalamus cells, we did not detect enhancer activity in the mouse adult hypothalamus. This could be due to several factors: 1) Established cell lines, in particular those from various neuronal cell types, can lose their tissue-specific properties; 2) Enhancer assays test sequences outside their genomic context; 3) The *trans* environment in the cells could allow activity of sequences that are not necessarily active in these cells; 4) *In vitro* luciferase assays test the ability of a sequence to turn on enhancer activity, but not the spatiotemporal activity of the sequence; 5) Due to cost and time limitations, we were only able to sample a few

time points in the mouse. It could be that OCE7 functions as a hypothalamus enhancer in different time points. Future mouse enhancer characterization of OCE7 and the other two functional OCEs could determine whether they drive enhancer activity in the hypothalamus or other tissues at different time points. In addition, there could be several other OCEs that were missed in our genomic annotation. Other technologies such as massively parallel reporter assays (MPRAs) could be used to assess all possible cCREs in the *OXTR* TAD boundary³¹.

Despite the caveats of our study, we found a 2 kb sequence in the intron of *OXTR* (OCE7) to function as an olfactory bulb enhancer and identified two other cCREs that demonstrate enhancer activity *in vitro*. Our study contributes to the growing body of literature annotating the factors that dictate *OXTR* expression. Despite being a highly conserved gene, *OXTR* expression in the brain and the periphery varies across species. cCREs might be one of the key regulators of *OXTR* expression that allows for its diverse species expression. Characterizing the elements that regulate *OXTR* expression provides a deeper understanding of its regulatory network and how variation in these sequences could dictate species specific expression of *OXTR* and how it might be associated with behavioral phenotypes in humans.

Figures and Tables

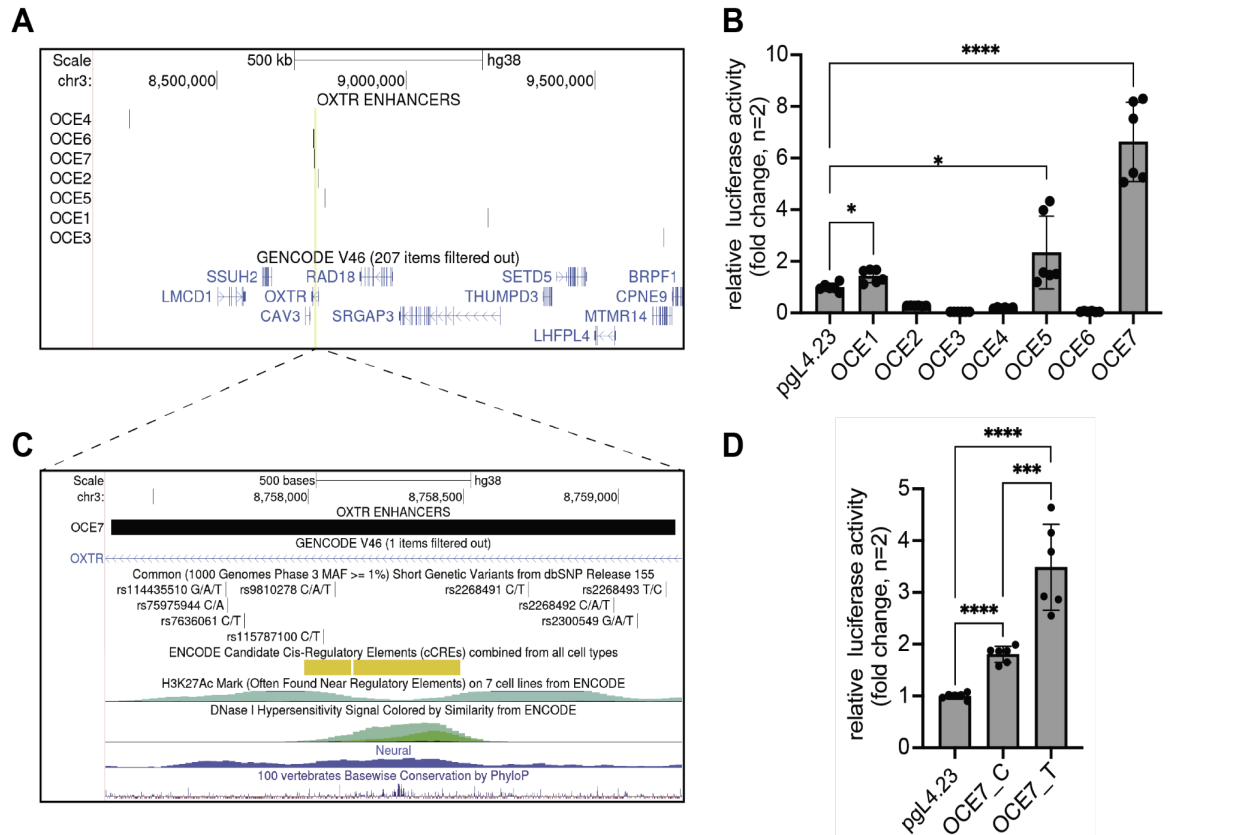


Figure 3.1: *OXTR* locus and OCE luciferase assays in hypothalamus cells.

(A) UCSC Genome Browser snapshot of the *OXTR* locus and *OXTR* candidate enhancer elements tested via luciferase assay. **(B)** Luciferase reporter assay of *OXTR* cCREs. Relative firefly/Renilla luciferase values are normalized to the empty vector following transient transfection in mHypoA-Pomc cells. pGL4.23 (empty vector) and candidate enhancers. Bars represent standard error for relative luciferase values from two biological replicates of triplicate measurements. Student's t-test * $p \leq 0.05$ and **** $p < 0.0001$. **(C)** Magnified view of OCE7 region (chr3:8757366-8759184; hg38) showing the following tracks: Short Genetic Variants from dbSNP release 155, GENE V46, ENCODE cCREs, H3K27ac ChIP-seq data from ENCODE, DNA Hypersensitivity data from ENCODE and DNA Hypersensitivity data from PsychENCODE and conservation across species. **(D)** Effect of rs2268491 reference and alternate allele on OCE7 enhancer activity measured via luciferase assay. Relative firefly/Renilla luciferase values are normalized to the empty vector following transient transfection in mHypoA-Pomc cells. pGL4.23 (empty vector), OCE7_C (OCE7 sequence with the reference C allele at rs2268491), and OCE7_T (OCE7 sequence with the alternate T allele at rs2268491). Bars represent standard error for relative luciferase values from two biological replicates of triplicate measurements. Student's t-test *** $p = 0.0006$ and **** $p < 0.0001$

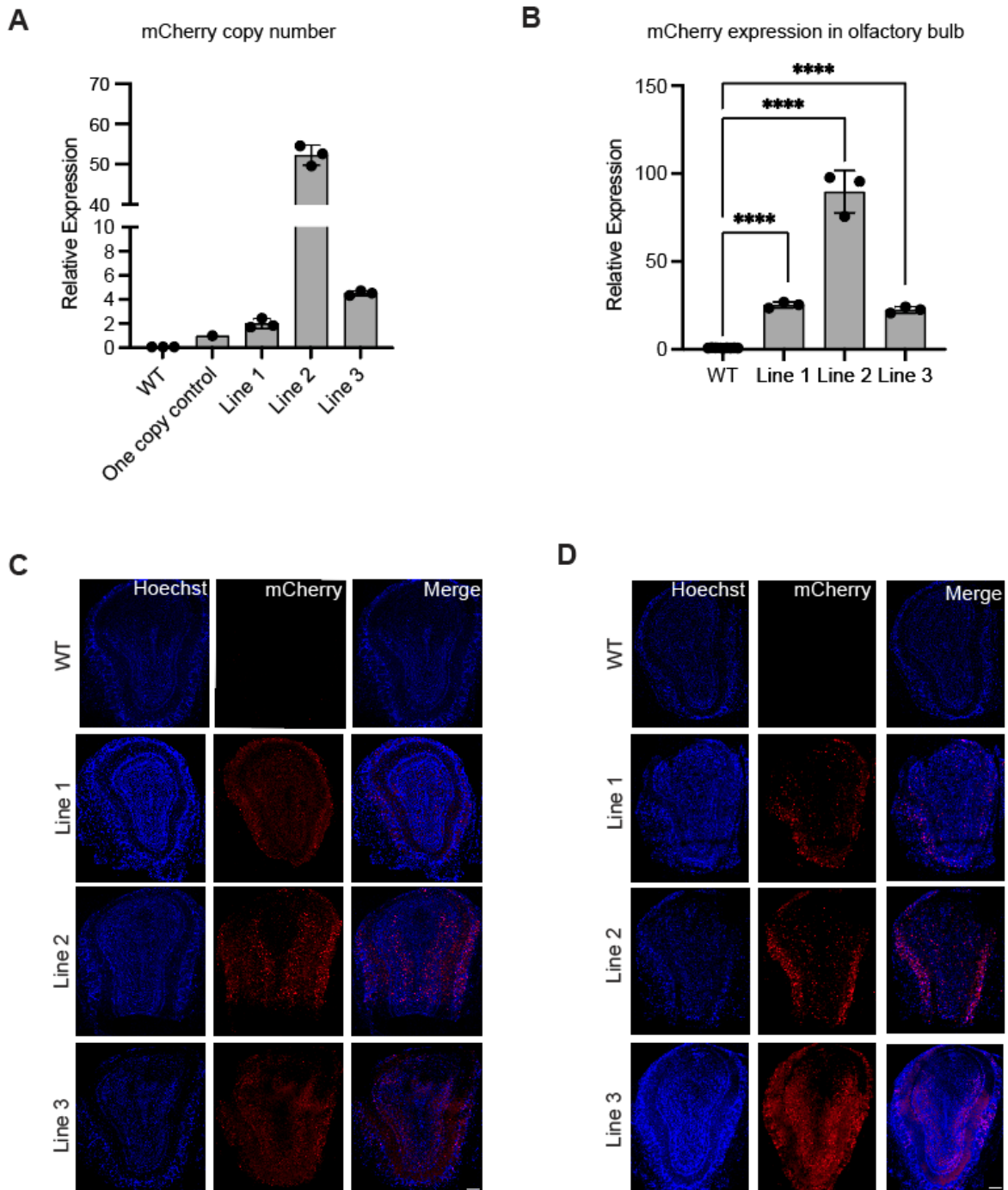
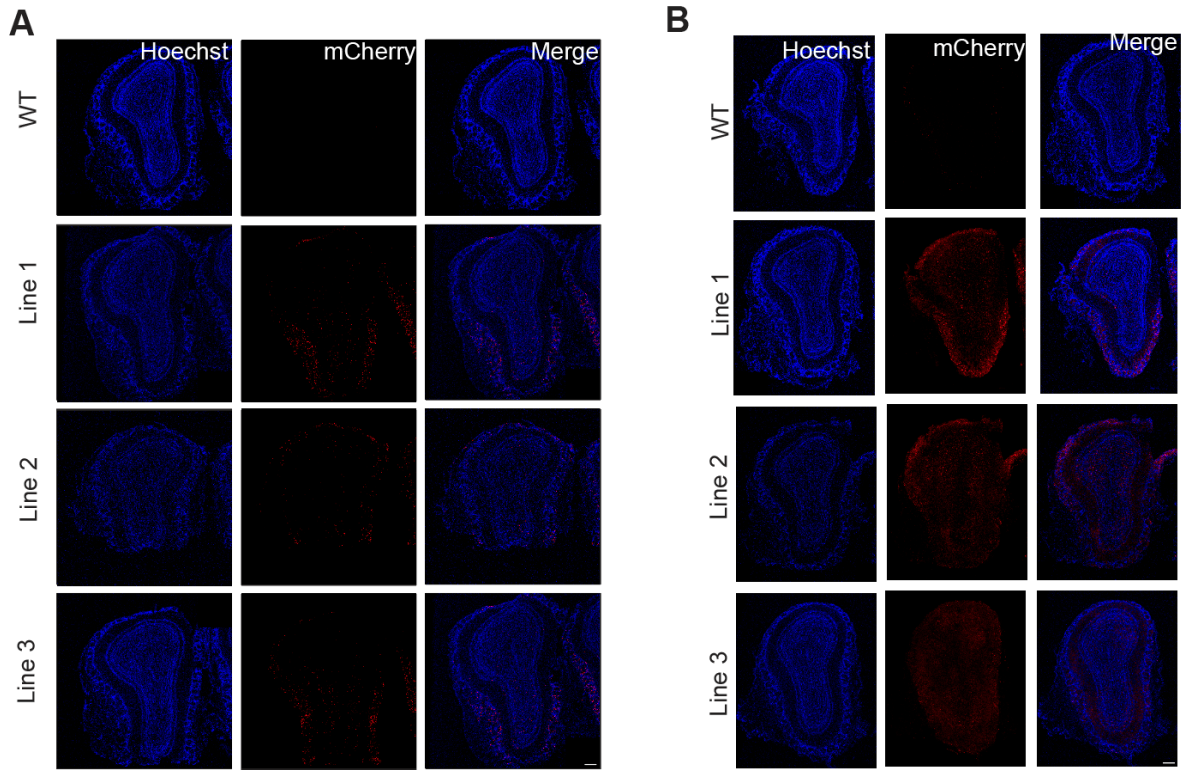


Figure 3.2: OCE7 mouse enhancer transgenic assay shows enhancer activity in the mouse olfactory bulb at postnatal day 28.

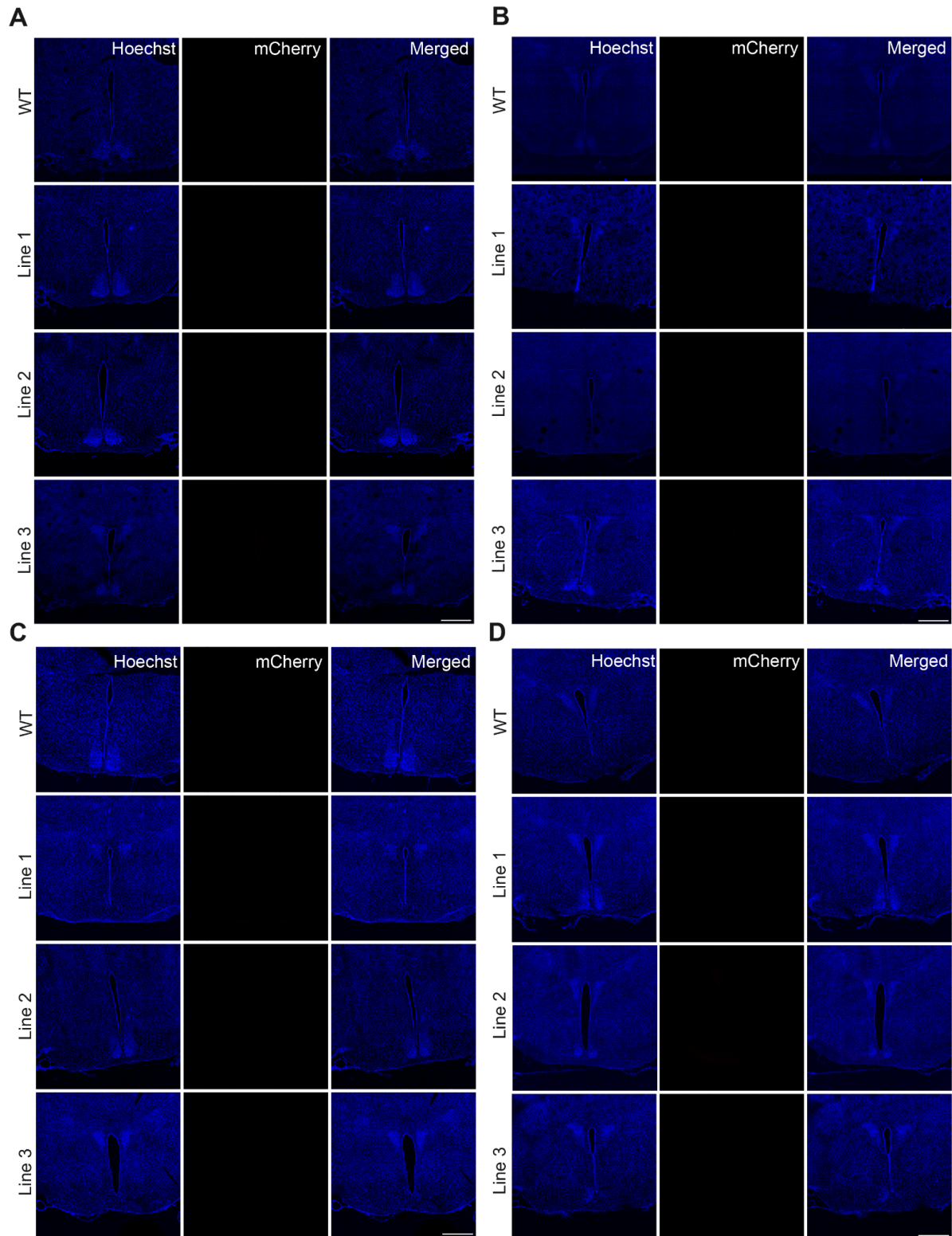
(A) qPCR for mCherry in the three OCE7 transgenic lines. qPCR was performed on DNA extracted from mouse tissue using *mCherry* and *Gapdh* (housekeeping gene) primers. We used wild-type (WT) genomic DNA and DNA from a mouse with one copy (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
of mCherry (one copy control) integrated into the genome as controls to determine transgene copy number. Bars represent standard error for the number of mCherry copies of three mice per line. **(B)** mCherry expression in OCE7 mouse enhancer transgenics lines via qRT-PCR in mouse olfactory bulb tissue at P28. RNA was extracted from the mouse tissue and qPCR was done using *mCherry* and *Gapdh* primers. Bars represent the standard error for the relative expression of mCherry for three mice per line. **(C-D)** Immunostaining showing mCherry expression in the olfactory bulb. Coronal section (35 μ M) of male **(C)** and female **(D)** mouse brains at P28 with Hoechst staining (blue) and anti-mCherry/A488 (recolorized as red).



Supplementary Figure 3.1: OCE7 mouse enhancer transgenic assay shows enhancer activity in the mouse olfactory bulb at postnatal day 56.

(A-B) Immunostaining showing mCherry expression in the olfactory bulb. Coronal section (35 μ M) of male **(A)** and female **(B)** mouse brains at P56 with Hoechst staining (blue) and anti-mCherry/A488 (recolorized as red). Scale bar = .2 mm.



Supplementary Figure 3.2: OCE7 mouse enhancer transgenic assay shows no enhancer activity in the mouse hypothalamus at postnatal day 28 and day 56.
 (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(A-B) Immunostaining showing mCherry expression in the hypothalamus. Coronal section (35 μM) of male **(A)** and female **(B)** mouse brains at P28 with Hoechst staining (blue) and anti-mCherry/A488 (recolorized as red). Scale bar = 0.5 mm. **(C-D)**

Immunostaining showing mCherry expression in the hypothalamus. Coronal section (35 μM) of male **(C)** and female **(D)** mouse brains at P56 with Hoechst staining (blue) and anti-mCherry/A488 (recolorized as red). Scale bar = 0.5 mm.

Supplementary Table 3.1: OCEs genomic coordinates and primers for cloning OCEs.

OXTR Candidate Enhancer	Coordinates (hg38)	Forward primer	Reverse primer
OCE1	chr3:9216878-9216942	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGCTAGCCCAGTGCTC TCAGAAA	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTGGCAGCAATGTTT TCATTTAG
OCE2	chr3:8767848-8768429	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGTGGGATCTTGGCC TTGGAGA	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTATTCCCGCTCATT TGCAGTGG
OCE3	chr3:9681951-9682518	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGCTTGTGACCAGCTG ATCTTCC	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTAGGAAGCCAGTAA AGGTAAACG
OCE4	chr3:8266807-8267046	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGGCTGTGTACAGAGC AAAGTTCC	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTCCAATCATTCTT CCATGTCA
OCE5	chr3:8784846-8785435	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGTCCCTTGTCTGGA ATCTGGGA	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTGGTTAGTGATTA GTTTCACTC
OCE6	chr3:8754836-8756640	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGTAATTGGTGACCT GTTGGA	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTCATCCACAGCGTC AAAAATG
OCE7	chr3:8757366-8759184	TGGCCGGTACCTGAGCTCGCTA GCCTCGAGGACTCCCAATCCCA GAACAA	TGTCTAAGCTTGGCCGCCGAG GCCAGATCTCCTGGCTTGTCTT CTTCCAG

Supplementary Table 3.2: Linkage disequilibrium analysis of SNPs around OCE7.

RS_number	rs237875	rs237878	rs17297803	rs237879	rs237880	rs11706648	rs237887	rs2268490	rs237888	rs9840864	rs4686301	rs2268491	rs2268492	rs2268493	rs2254298	rs53576
rs237875	1	0.424	0.17	0.034	0.385	0.026	0.084	0.033	0.029	0.015	0.014	0.031	0.036	0.036	0.031	0.005
rs237878	0.424	1	0.071	0.081	0.931	0.05	0.042	0.019	0.068	0.002	0.038	0.018	0.001	0.001	0.017	0.001
rs17297803	0.17	0.071	1	0.006	0.074	0.012	0.02	0.012	0.005	0.014	0.013	0.01	0.035	0.035	0.01	0.006
rs237879	0.034	0.081	0.006	1	0.077	0.002	0.002	0.001	0.08	0	0.006	0	0.001	0.001	0	0.01
rs237880	0.385	0.931	0.074	0.077	1	0.046	0.035	0.019	0.063	0.003	0.035	0.018	0	0	0.017	0
rs11706648	0.026	0.05	0.012	0.002	0.046	1	0.392	0.074	0.03	0.132	0.884	0.045	0.35	0.35	0.047	0.067
rs237887	0.084	0.042	0.02	0.002	0.035	0.392	1	0.192	0.054	0.025	0.35	0.12	0.243	0.243	0.123	0.142
rs2268490	0.033	0.019	0.012	0.001	0.019	0.074	0.192	1	0.01	0.51	0.055	0.67	0.071	0.071	0.67	0.028
rs237888	0.029	0.068	0.005	0.08	0.063	0.03	0.054	0.01	1	0.052	0.028	0.008	0.027	0.027	0.008	0.009
rs9840864	0.015	0.002	0.014	0	0.003	0.132	0.025	0.51	0.052	1	0.128	0.387	0.124	0.124	0.386	0.011
rs4686301	0.014	0.038	0.013	0.006	0.035	0.884	0.35	0.055	0.028	0.128	1	0.043	0.297	0.297	0.044	0.068
rs2268491	0.031	0.018	0.01	0	0.018	0.045	0.12	0.67	0.008	0.387	0.043	1	0.053	0.053	0.98	0.021
rs2268492	0.036	0.001	0.035	0.001	0	0.35	0.243	0.071	0.027	0.124	0.297	0.053	1	1	0.054	0.072
rs2268493	0.036	0.001	0.035	0.001	0	0.35	0.243	0.071	0.027	0.124	0.297	0.053	1	1	0.054	0.072
rs2254298	0.031	0.017	0.01	0	0.017	0.047	0.123	0.67	0.008	0.386	0.044	0.98	0.054	0.054	1	0.02
rs53576	0.005	0.001	0.006	0.01	0	0.067	0.142	0.028	0.009	0.011	0.068	0.021	0.072	0.072	0.02	1

Supplementary Table 3.3: Primer sequences.

Primer	Forward	Reverse
OXTR-HSP68	CGAGGTCGACGGTATCGATAGCGGCCG CTTGAGATCAAGAACGGTGGA	TTGTTCTGGGATTGGGAGTCTGATATC GAATTCCTGCAGC
mCherry	ACTACGACGCTGAGGTCAAG	CTCGTTGTGGGAGGTGATGT
Gapdh	CATCACTGCCACCCAGAAGACTG	ATGCCAGTGAGCTTCCCGTTCAG
OCE7_SDM	AAGAGCCAAACGGGCGGGCTA	CACCCAGACCTTGCACTAC

Materials and Methods

Candidate enhancer sequence selection

Candidate enhancer sequences were selected from mouse and human hypothalamus scRNA-seq/scATAC-seq generated in our lab²⁵. Peaks correlated to *OXTR* gene expression were found using the LoupeBrowser Feature Linkage table³².

Luciferase Assays

Candidate enhancer sequences were PCR amplified from human genomic DNA (Takara Bio, 636401) using specific primers (Table S1). Sequences were cloned into the pGL4.23 plasmid (Promega) upstream of a minimal promoter and firefly luciferase reporter using the NEB HiFi DNA Assembly Cloning kit (NEB) following the manufacturer's protocol. The vectors with the candidate enhancer sequences were verified via Sanger sequencing. mHypoA-POMC/GFP-1 cells were seeded in a 24 well plate with 5×10^4 cells per well 24 hours before transfection. Using the X-treme Gene HP Transfection Reagent (Roche) 450 ng of the OCE plasmids were individually transfected along with 50 ng of pGL4.74 (Promega), to control for transfection efficiency. As a negative control, we used an empty pGL4.23 vector and an SV40 enhancer (pGL4.13; Promega) as a positive control. The DNA:X-tremeGENE ratio was 1:3. The firefly luciferase and Renilla activity were measured 48 hours post-transfection following the Promega Dual-Luciferase Assay protocol on a GloMax Explorer Multimode Microplate Reader (Promega). Renilla activity was used to normalize the firefly luciferase activity. The luciferase reporter assays were performed in duplicates and on two different days to obtain two biological replicates. The expression value of the empty vector (pGL4.23) was arbitrarily set at 1.0, and statistical differences between vectors

containing candidate enhancer sequences were determined using a two-sided unpaired t-test. For the positive control, pGL4.13 (Promega), the relative luciferase activity was determined to be 1191.47 and standard error of 348.36.

Site-Directed Mutagenesis

The OCE7_C was generated using the Q5 Site-Directed Mutagenesis Kit (NEB) following the manufacturer's protocol. First, back-to-back primers were designed using NEBase Changer³³. Then, the exponential amplification was conducted following the manufacturer's protocol using the OCE7_SDM primer set (Table S3) and the pGL4.23 plasmid with OCE7 cloned upstream of the minimal promoter as a template. The resulting PCR product underwent a KLD reaction following the manufacturer's protocol and the resulting plasmid was transformed using chemically-competent cells and amplified using the QIAGEN Plasmid Midi Kit (QIAGEN). The resulting vector was verified via Sanger sequencing.

Generation of transgenic mice

For the mouse transgenic enhancer assays, we used an Hsp68-mCherry (hCR) plasmid, a gift from Drs. Len Pennacchio, Axel Visel and Dianne Dickel at the Lawrence Berkley National Laboratory³⁵.

The OCE7 (chr3:8757366-8759184; hg38) was amplified by the OXTR-HSP68 primer set (Table S3) with human genomic DNA (Takara Bio, 636401) and cloned into the Hsp68-mCherry plasmid after digestion with *KpnI*. The resulting plasmid was digested with *Sall*, and the DNA fragment was released from the backbone vector and used for pronuclear injection, which was performed by the Transgenic Gene Targeting Core at the Gladstone Institute. All mouse work was approved by the UCSF IACUC protocol

number AN197608 and was conducted in accordance with AALAC and NIH guidelines. The C57BL/6NHsd mouse strain (ENVIGO; 044) was used.

Quantitative RT-PCR

For the mCherry copy number determination, DNA was collected from mouse tail clip using Trizol TRIzol (Thermo Fisher Scientific; 15596026) and qPCR was performed using SsoFast EvaGreen supermix (Bio Rad; 1725205) on QuantStudio™ 6 Flex Real-Time PCR System (Applied Biosystems; 4485691) using *mCherry* and *Gapdh* primers (Table S3). To normalize the mCherry copy number of the transgenic lines, we obtained DNA from the knockin reporter mouse line (*I33^{mCherry/+}*), which contains one copy of mCherry, as a kind gift from Dr. Anna Molofsky³⁴. The copy number was determined by using the $\Delta\Delta$ CT method compared to the knockin reporter mouse line (*I33^{mCherry/+}*) and normalized to *Gapdh* as a housekeeping gene. For the quantification of mCherry expression in the olfactory bulb, RNA was collected from the mouse olfactory bulb using TRIzol (Thermo Fisher Scientific; 15596026) and cDNA was generated from total RNA using ReverTra Ace qPCR-RT master mix with genomic DNA (gDNA) remover (Toyobo; FSQ-301). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad; 1725205) on QuantStudio™ 6 Flex Real-Time PCR System (Applied Biosystems; 4485691) using mRNA *mCherry* and *Gapdh* primers (Table S3). The relative expression of mCherry was analyzed using the $\Delta\Delta$ CT method compared to wildtype and normalized to *Gapdh* as a housekeeping gene.

Immunostaining

Female and male mice were evaluated from each mouse line at each post-developmental time point (P28 and P56). Mice were anesthetized intraperitoneally with tribromoethanol (avertin) and transcardially perfused with 10 mL of Dulbecco's Phosphate-Buffered Saline (DPBS) followed by 10 mL of 4% paraformaldehyde (PFA) in DPBS. Brains were removed and post-fixed for 24 hours in 4% PFA. Brains were then equilibrated in 30% sucrose in DPBS for 48 hours and then embedded and sectioned coronally (35 μ m) on a cryostat. Slides with coronal brain sections that had been stored at -80 °C were permeabilized for 5 minutes at room temperature with 0.5% TX-100/DPBS and blocked with 5% goat serum/0.2% Tween/DPBS for 1 hour at room temperature. Coronal sections were then incubated overnight at 4 °C with rabbit monoclonal anti-mCherry (Cell Signaling Technologies; 43590) as the primary antibody at a dilution of 1:1000. After overnight incubation with primary antibody, coronal sections were incubated at room temperature for 1 hour with anti-rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488 (Invitrogen; A11008) as the secondary antibody at a dilution of 1:2000 and Hoechst at a dilution of 1:20000. A coverslip was placed on the slides along with Prolong Antifade mounting media. Images were captured with Zeiss microscope (Zeiss LSM 700 Confocal).

References

1. Jurek, B. & Neumann, I. D. The Oxytocin Receptor: From Intracellular Signaling to Behavior. *Physiol. Rev.* **98**, 1805–1908 (2018).
2. Meyer-Lindenberg, A., Domes, G., Kirsch, P. & Heinrichs, M. Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nat. Rev. Neurosci.* **12**, 524–538 (2011).
3. Gregory, S. G. *et al.* Genomic and epigenetic evidence for oxytocin receptor deficiency in autism. *BMC Med.* **7**, 62 (2009).
4. Gouin, J. P. *et al.* Associations among oxytocin receptor gene (OXTR) DNA methylation in adulthood, exposure to early life adversity, and childhood trajectories of anxiousness. *Sci. Rep.* **7**, 7446 (2017).
5. Ziegler, C. *et al.* Oxytocin receptor gene methylation: converging multilevel evidence for a role in social anxiety. *Neuropsychopharmacology* **40**, 1528–1538 (2015).
6. Maud, C., Ryan, J., McIntosh, J. E. & Olsson, C. A. The role of oxytocin receptor gene (OXTR) DNA methylation (DNAm) in human social and emotional functioning: a systematic narrative review. *BMC Psychiatry* **18**, 154 (2018).
7. Kraaijenvanger, E. J. *et al.* Epigenetic variability in the human oxytocin receptor (OXTR) gene: A possible pathway from early life experiences to psychopathologies. *Neurosci. Biobehav. Rev.* **96**, 127–142 (2019).
8. Skuse, D. H. *et al.* Common polymorphism in the oxytocin receptor gene (OXTR) is associated with human social recognition skills. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1987–1992 (2014).

9. Rodrigues, S. M., Saslow, L. R., Garcia, N., John, O. P. & Keltner, D. Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21437–21441 (2009).
10. Huetter, F. K. *et al.* Association of a Common Oxytocin Receptor Gene Polymorphism with Self-Reported ‘Empathic Concern’ in a Large Population of Healthy Volunteers. *PLoS One* **11**, e0160059 (2016).
11. Cattaneo, Z. *et al.* Congenital prosopagnosia is associated with a genetic variation in the oxytocin receptor (OXTR) gene: An exploratory study. *Neuroscience* **339**, 162–173 (2016).
12. Tost, H. *et al.* A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 13936–13941 (2010).
13. Choi, D., Tsuchiya, K. J. & Takei, N. Interaction effect of oxytocin receptor (OXTR) rs53576 genotype and maternal postpartum depression on child behavioural problems. *Sci. Rep.* **9**, 7685 (2019).
14. Ferguson, J. N. *et al.* Social amnesia in mice lacking the oxytocin gene. *Nat. Genet.* **25**, 284–288 (2000).
15. Takayanagi, Y. *et al.* Pervasive social deficits, but normal parturition, in oxytocin receptor-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 16096–16101 (2005).
16. Sala, M. *et al.* Mice heterozygous for the oxytocin receptor gene (Oxtr(+/-)) show impaired social behaviour but not increased aggression or cognitive inflexibility: evidence of a selective haploinsufficiency gene effect. *J. Neuroendocrinol.* **25**, 107–118 (2013).

17. Matharu, N. & Ahituv, N. Modulating gene regulation to treat genetic disorders. *Nat. Rev. Drug Discov.* **19**, 757–775 (2020).
18. Theofanopoulou, C., Andirkó, A., Boeckx, C. & Jarvis, E. D. Oxytocin and vasotocin receptor variation and the evolution of human prosociality. *Compr Psychoneuroendocrinol* **11**, 100139 (2022).
19. Mizumoto, Y., Kimura, T. & Ivell, R. A genomic element within the third intron of the human oxytocin receptor gene may be involved in transcriptional suppression. *Mol. Cell. Endocrinol.* **135**, 129–138 (1997).
20. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
21. Zhang, Q. *et al.* Distal regulatory sequences contribute to diversity in brain oxytocin receptor expression patterns and social behavior. *bioRxiv* 2022.12.01.518660 (2023) doi:10.1101/2022.12.01.518660.
22. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
23. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
24. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).
25. Nguyen, H. P. *et al.* Integrative single-cell characterization of hypothalamus sex-differential and obesity-associated genes and regulatory elements. *bioRxiv* 2022.11.06.515311 (2022) doi:10.1101/2022.11.06.515311.

26. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmüller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**, 1334–1336 (2015).
27. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491–499.e15 (2018).
28. Newmaster, K. T. *et al.* Quantitative cellular-resolution map of the oxytocin receptor in postnatally developing mouse brains. *Nat. Commun.* **11**, 1885 (2020).
29. Quintana, D. S. *et al.* Oxytocin pathway gene networks in the human brain. *Nat. Commun.* **10**, 668 (2019).
30. Rokicki, J. *et al.* Oxytocin receptor expression patterns in the human brain across development. *Neuropsychopharmacology* **47**, 1550–1560 (2022).
31. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
32. Loupe Browser - Official 10x Genomics Support. *10x Genomics*
<https://www.10xgenomics.com/support/software/loupe-browser/latest>.
33. NEBaseChanger_V1. <https://nebasechangerv1.neb.com/>.
34. Nguyen, P. T. *et al.* Microglial Remodeling of the Extracellular Matrix Promotes Synapse Plasticity. *Cell* **182**, 388–403.e15 (2020).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

Dianne Laboy Cintron
40045800DB18435... Author Signature

8/28/2024
Date