Title
Characterizing the epigenomic landscape of the developing human cortex at single-cell resolution

Author
Ziffra, Ryan

Publication Date
2021

Peer reviewed|Thesis/dissertation

Characterizing the epigenomic landscape of the developing human cortex at single-cell resolution

by
Ryan Ziffra

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:
44A19A16E413450...
John Rubenstein
Chair

DocuSigned by:
Benoit Bruneau
Benoit Bruneau

DocuSigned by:
Tomasz Nowakowski
Tomasz Nowakowski

DocuSigned by:
Nadav Ahituv
77BA96E7DAE34F4...
Nadav Ahituv

Committee Members

This work is dedicated to my wife Amanda and son Jay, who provide me with the support and motivation to persevere through the most challenging times.

ACKNOWLEDGEMENTS

The journey to complete my doctoral work has been an unexpected adventure spanning the full range of human emotion. Throughout my time at UCSF, there have been moments when I have felt productive, burned-out, exhilarated, defeated, joyful, and depressed. In hindsight, it is clear to me that, despite the ups and downs, the experience has overall been extremely valuable and fundamentally changed me as a person for the better. It is also clear that, without the support of a whole team of loving and caring individuals behind me, there is no way I would have been able to make it through to the end.

First and foremost, I'd like to thank my advisors Tomasz Nowakowski and Nadav Ahituv, who have always been completely supportive of my goals, even when they didn't necessarily align perfectly with their own. I will always remember the first time I met with Tom, who had just established his lab at UCSF that year, when he walked me through all of the project ideas he had for the lab. His passion for scientific inquiry, the wonders of the human brain, and the promise of novel molecular technologies was (and still is) highly contagious. I have always envied Tom for his passion and utter dedication to his craft. I know if I walk away from my experience as a graduate student in his lab with 1/10th of the passion he has for his work that it will be sufficient to motivate me for my entire career. Likewise, I have always been blown away by the breadth of work that occurs in Nadav's lab. I still have no idea how he keeps track of it all. Lab meeting could be about bat wing development one day, frog parasites the next, and genetic therapies for neurodevelopmental diseases the following. It has been such an exciting environment to

be a part of and I am so grateful to have had the experience. Throughout it all, Tom and Nadav has always been there to provide me with the mentorship, feedback, and support I needed to get through the inevitable difficulties of graduate training and the leeway to pursue career development opportunities outside of the lab. I am forever grateful to you both.

I would also like to thank a handful of other mentors at UCSF who have been instrumental in my development as a scientist. My thesis committee members John Rubenstein and Benoit Bruneau, my qualifying exam committee members Stephan Sanders, Vijay Ramani, and Katie Pollard, and other close mentors Alex Pollen and Yin Shen. Thank you all for taking the time to make yourselves available to me throughout the years and helping me to grow as a scientist. I'd also like to thank Ryan Delgado and Taka Inoue, former postdocs in the Nowakowski and Ahituv labs respectively, who spent a tremendous amount of time and effort training me as a rotation student.

I am so lucky to have been a part of two lab families populated with such amazing and collaborative people. Thank you to all my lab mates in the Ahituv and Nowakowski labs who taught me so much, endured all the practice talks, and didn't laugh at me (publicly) when I didn't know how to turn on the confocal. You all provided much needed reprieve from what could sometimes be a grueling and monotonous experience. I will forever cherish the pre-pandemic lab lunches we had together in the Nowakowski lab and the holiday & pool parties with the Ahituv lab.

Thank you to the BMS program administrators Demian Sainz and Ned Molyneaux for somehow making graduate school a navigable path and program directors Mark Ansel and Anita Sil for providing me with the opportunity and structure to develop into the scientist I am becoming. Thank you to Autism Speaks for supporting me these past two years as a predoctoral fellow. I am hopeful that my work will, at least in some small way, contribute to the efforts to alleviate the suffering of those afflicted by debilitating symptoms.

I'd also like to acknowledge a few others who were instrumental in my path to joining UCSF. Thank you to Carole Ober and Marcelo Nobrega at the University of Chicago for taking a chance on me and bringing me on as a tech with little to know relevant research experience. Thank you to Scott Kreher at Dominican University for opening my eyes to the world of genetics and molecular biology. Scott is the first person who truly inspired me to pursue biological research and provided me with primary research opportunities. Thank you for (indirectly) convincing me to pursue graduate school instead of medical school.

Lastly, and most importantly, I'd like to thank my family. To my parents Kevin and Dahlia, thank you for never giving up on me, even when things seemed like they might be hopeless. Thank you for sticking by me and hearing me out when I changed my mind over and over again about my career goals. And thank you for always being my biggest supporters through all of life's challenges. Thank you to my brother Nick for helping me to survive those long family road trips and holiday parties. To my wife Amanda, I could

not have asked for a more supportive, caring, and understanding partner. Thank you for allowing me this period of development by putting up with all the late nights of work and supporting our family during this time. You are my best friend and I love you endlessly. To my son Jay, in the brief one-year span that you have been in my life, you have become my raison d'être. Thank you for bringing a level of purpose and motivation to my life that I never imagined was possible. I promise to do my best to use the experience I have gained during this time to make the world a better place for you.

In every scientist's career, there is that one moment when you discover something novel and, for that one precious moment in time, you are the only person in the world with that piece of knowledge. The prospect of attaining such uniquely personal knowledge is one of the motivators that excited me about scientific research and drove me to pursue a PhD. Coming out the other end of my graduate training, it is not the attainment of knowledge that drives me, but rather, it is the responsibility to share that knowledge with the rest of the world to advance our collective understanding of the truth. Thank you to all the scientists who came before me and chose to share those personal truths with the world.

CONTRIBUTIONS

All work described in this dissertation was carried out under the direct supervision of Drs. Tomasz Nowakowski and Nadav Ahituv. Chapter 2 is adapted from a research article that has been peer-reviewed, accepted, and is in press at Nature, of which Ryan is the sole first author. Additional author contributions to this article are described in the Contributions section of Chapter 2.

Citation for Chapter 2 manuscript:

Ziffra RS, Kim CN, Ross JM, Wilfert A, Turner TN, Haeussler M, Casella AM, Przytycki PF, Keough KC, Shin D, Bogdanoff D, Kreimer A, Pollard KS, Ament SA, Eichler EE, Ahituv N, Nowakowski TJ. (in press). Single cell epigenomics reveals mechanisms of human cortical development. *Nature.*

Characterizing the epigenomic landscape of the developing human cortex at single-cell resolution

Ryan Ziffra

ABSTRACT

During mammalian development, chromatin state differences coincide with cellular differentiation and reflect changes in the gene regulatory landscape. In the developing brain, cell fate specification and topographic identity play important roles in defining cell identity and confer selective vulnerabilities to neurodevelopmental disorders. To identify cell type specific chromatin accessibility patterns in the developing human brain, we used a single cell assay for transposase accessibility by sequencing (scATAC-seq) in primary human forebrain tissue samples. We applied unbiased analyses to identify genomic loci that undergo extensive cell type- and brain region-specific changes in accessibility during neurogenesis and an integrative analysis to predict cell type specific candidate regulatory elements. We found that cerebral organoids recapitulate most putative cell type-specific enhancer accessibility patterns but lack many cell type specific open chromatin regions found *in vivo*. Systematic comparison of chromatin accessibility across brain regions revealed an unexpected diversity among neural progenitor cells in the cerebral cortex and implicate retinoic acid signaling in the specification of prefrontal cortex neuronal lineage identity. Together, our results reveal the important contribution of chromatin state to the emerging patterns of cell type diversity and cell fate specification and provide a blueprint for evaluating the fidelity and robustness of cerebral organoids as a model for cortical development.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

The cerebral cortex is the largest region of the human brain, representing roughly 80% of the total brain mass, and is responsible for higher order cognitive functions, such as language, decision making, and sensory perception[1]. The cortex underwent great expansion during human evolution, and its larger size in humans relative to other mammals is thought to be responsible for our extraordinary cognitive abilities[2]. Since the times of Santiago Ramon y Cajal, it has been appreciated that the cerebral cortex contains an astonishing diversity of cell types distributed across dozens of anatomically distinct areas[3,4]. These distinct cortical areas can be distinguished on the basis of sharply bordered patterns of cytoarchitecture, including differences in cell density, lamination, and morphology, and often correspond with functional specialization[5–10]. Despite the vast body of research establishing and characterizing the differences between cortical areas in adult brains, few of these differences are apparent during development, and the specific mechanisms and timing by which they attain their distinct anatomical and functional identities is still largely unknown[4].

Hypotheses of cortical development

During embryonic development, the neuroepithelial tissue that gives rise to the majority of cortical cells originates as a homogeneous layer of neuroepithelial stem cells. Over the past few decades, neurobiologists have circled around two main hypotheses which explain how this initially unform layer of progenitors gives rise to the vast diversity of cell types and functionally specialized areas found in the adult human cortex. The “protomap”

hypothesis suggests that the cortical progenitors are pre-programmed to give rise to area-specific cell types based on intrinsic developmental programs[11–13] while the “protocortex” hypothesis suggests that the cortex is initially homogenous and takes on area-specific architecture following differential extrinsic signals, such as input from the thalamus[14–16].

Early forebrain patterning

Much has already been established as relates to the patterning of the forebrain during prenatal development. Morphogen gradients of FGF, WNT, BMP, SHH, and TGFα are established early on and induce the expression of patterning transcription factors (TFs), including Pax6, Sp8, Emx2, and COUP-TFI[3,17]. These patterning TFs are expressed in gradients along the anterior-posterior and lateral-medial axes of the developing forebrain and induce the expression of additional TFs that initiate area-specific developmental programs controlling the cell fate specification of molecularly distinct neuronal subtypes. Due to the fact that these TFs are expressed in graded patterns, it is clear that they must work in concert with additional mechanisms in order to give rise to the sharply bordered cortical areas found in the fully developed adult cortex.

The Epigenetic Landscape

Over 60 years ago, Conrad Waddington introduced the concept of an epigenetic landscape to account for the emergence of distinct cell fates[18]. In particular, chromatin state defines the functional architecture of the genome by modulating the accessibility of gene regulatory elements (GREs), such as enhancers and promoters, which serve as binding sites for transcriptional regulators that enable the assembly of the transcriptional

machinery. Together with the expression of unique combinations of graded TFs, chromatin state is believed to represent the *cis*-regulatory 'vocabulary' of gene expression, which is a fundamental determinant of cell identity[19,20]. However, studies of chromatin state in developing brain have been limited because established methods for discovering GREs, such as the Assay for Transposase-Accessible Chromatin with Sequencing (ATAC-seq)[21] or Chromatin Immunoprecipitation with Sequencing (ChIP-seq)[22], lacked cellular resolution. To gain insight into cell type specific patterns of chromatin state changes during development, few studies have been able to enrich for, or isolate, broad cell classes and perform measurements of chromatin state or other epigenetic assays[23–29] and revealed changes in chromatin state broadly correlated with regional patterning and neuronal differentiation. For example, many mouse brain enhancers, are developmentally regulated[27] and exhibit sharp regional boundaries of activity in the subpallium (which gives rise to the basal ganglia and cortical interneurons)[19,20]. These sharply bordered patterns of enhancer activity suggest that enhancers may play an important role in the specification of anatomically defined areas of the cortex. However, the cell type specific nature of many enhancers makes them difficult to study in the context of heterogeneous tissues, such as the brain, using bulk assays that provide an average signal of the sample. Future studies of the epigenomic landscape of the developing brain will benefit greatly from approaches that are able to profile the regulator landscape of discrete cell types.

Molecular diversity of the developing cortex

The diverse cell types of the cortex have historically been classified based on a handful of morphological, anatomical, and physiological features. However, during developmental stages, many of these features are indistinguishable across cortical areas. More recent studies, which have examined the molecular profiles of cells in the developing cortex, have begun to identify and characterize distinctions between cells in different cortical areas. One of the first studies to systematically assess molecular differences between cortical areas during development using exon arrays found hundreds of genes differentially expressed between rostral and caudal areas[30]. While bulk transcriptomic studies such as this made it clear that molecular differences between cells from different cortical areas were detectable, their ability to elucidate the expression programs that lead to the divergence of closely related cell types was limited by the nature of bulk assays, which provide an average profile of the sampled region. Recent innovations in single cell genomics, such as single cell mRNA sequencing (scRNA-seq), have enabled massively parallel profiling of thousands of molecular features in every cell, uncovering the remarkable molecular diversity of cell types previously considered homologous, such as excitatory neurons located in different areas of the cerebral cortex[31–34]. While single cell transcriptomic studies have been invaluable in redefining the molecular diversity of cell types in the cortex, they have not answered the question of how distinct gene expression patterns are established through the modulation of transcriptional regulators.

The promise of single cell epigenomics

Recently, the advances in single cell genomics have been extended to epigenomic assays, such as scATAC-seq[35,36], enabling the profiling of the epigenomic landscape at single cell resolution. Studies using these methods have revealed many cell type specific patterns of enhancer activity in the developing and adult mouse brain, as well as adult human brain[37–39]. GREs often contain binding sites for transcriptional regulators, enabling the discovery of regulatory programs that control the specification of specific cell types. To understand the regulatory landscape of the developing human cortex, it will be important to characterize GREs in their native context of developing human tissue directly, rather than drawing inferences from mouse and adult human tissues, as growing evidence suggests that human genome sequence evolution is accelerated in putative neurodevelopmental enhancer regions[40–42] and many enhancers exhibit transient patterns of activity[27].

Non-coding variation in neuropsychiatric disease

Mutations in non-coding genomic regions, as well as *de novo* loss of function mutations in chromatin modifiers have been implicated in a wide range of neurodevelopmental and psychiatric disorders, including schizophrenia[43] and autism spectrum disorder (ASD)[44–46]. Cellular-resolution datasets of chromatin state across developmental stages and differentiation states may provide an important link between these mutations and selective vulnerabilities among the diverse cell types of the developing human brain, as was seen with recent studies using single cell transcriptomic data[46,47]. A recent study was able to use scATAC-seq data from the adult human brain to fine map disease-associated

genomic loci to pinpoint the likely causal variants associated with Parkisnon's disease[48]. Similar datasets in the developing human brain may enable similar fine-mapping of neurodevelopmental disease-associated loci, shedding light on the cell type specific vulnerabilities and causal mechanisms by which common genetic variants may contribute the poorly understood heritability of neurodevelopmental diseases.

Cerebral organoids as a model for human cortical development

Due to the scarcity of primary human tissue, studies of human neural development critically require suitable *in vitro* models, such as cerebral organoids, a 3D cell culture system in which cells self-organize and recapitulate many macro-scale phenomena of neurodevelopment[49]. Previous studies emphasized the similarities between cerebral organoid cells and their *in vivo* counterparts using single cell transcriptomics[50,51] and bulk epigenomics[26,28,52]. However, others have noted the shortcomings of cerebral organoids, such as elevated signatures of cellular stress, which may compromise their integrity as a model of healthy developmental states[53]. Importantly, no studies have yet assessed the fidelity of cerebral organoids as a model for human epigenomic signatures with cellular resolution.

At the time this dissertation work was undertaken, there were no publicly available cellular resolution epigenomic datasets from the prenatal human cortex or cerebral organoid models. With this dissertation work, we sought to generate such datasets and use them to characterize the regulatory landscape of human cortical development, with specific

focus on understanding the mechanisms that underlie cell fate specification and arealization.

REFERENCES

1. Azevedo FAC, Carvalho LRB, Grinberg LT, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol*. 2009;513(5):532-541. doi:10.1002/CNE.21974
2. Geschwind DH, Rakic P. Cortical evolution: Judge the brain by its cover. *Neuron*. 2013;80(3):633-647. doi:10.1016/j.neuron.2013.10.045
3. Leary DDMO, Sahara S. Genetic regulation of arealization of the neocortex. 2008. doi:10.1016/j.conb.2008.05.011
4. Cadwell CR, Bhaduri A, Mostajo-Radji MA, Keefe MG, Nowakowski TJ. Development and Arealization of the Cerebral Cortex. *Neuron*. 2019;103(6):980-1004. doi:10.1016/j.neuron.2019.07.009
5. Brodmann K. *Vergleichende Lokalisationslehre Der Grosshirnrinde in Ihren Prinzipien Dargestellt Auf Grund Des Zellenbaues*. Barth; 1909.
6. von Economo CF, Koskinas GN. *Die Cytoarchitektonik Der Hirnrinde Des Erwachsenen Menschen*. J. Springer; 1925.
7. Defelipe J, González-Albo MC, Del Río MR, Elston GN. Distribution and patterns of connectivity of interneurons containing calbindin, calretinin, and parvalbumin in visual areas of the occipital and temporal lobes of the macaque monkey. *J Comp Neurol*. 1999;412(3):515-526.
8. Bayer SA, Altman J. *Neocortical Development*. Vol 1. Raven Press New York; 1991.
9. Elston GN. Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. *Cereb cortex*. 2003;13(11):1124-1138.

10. Hof PR, Nimchinsky EA. Regional distribution of neurofilament and calcium-binding proteins in the cingulate cortex of the macaque monkey. *Cereb Cortex*. 1992;2(6):456-467.
11. Bishop KM, Goudreau G, O'Leary DDM. Regulation of area identity in the mammalian neocortex by Emx2 and Pax6. *Science (80-)*. 2000;288(5464):344-349.
12. Rakic P. Specification of cerebral cortical areas. *Science (80-)*. 1988;241(4862):170-176.
13. Rubenstein JLR, Rakic P. Genetic control of cortical development. *Cereb Cortex*. 1999;9(6):521-523.
14. Van der Loos H, Woolsey TA. Somatosensory cortex: structural alterations following early injury to sense organs. *Science (80-)*. 1973;179(4071):395-398.
15. Creutzfeldt OD. Generality of the functional structure of the neocortex. *Naturwissenschaften*. 1977;64(10):507-517.
16. O'Leary DDM. Do cortical areas emerge from a protocortex? *Trends Neurosci*. 1989;12(10):400-406.
17. Rubenstein JL, Beachy PA. Patterning of the embryonic forebrain. *Curr Opin Neurobiol*. 1998;8(1):18-26. doi:10.1016/S0959-4388(98)80004-4
18. Waddington CH. *The Strategy of the Genes*. London: Routledge; 1957. doi:10.4324/9781315765471
19. Visel A, Taher L, Girgis H, et al. A high-resolution enhancer atlas of the developing telencephalon. *Cell*. 2013;152(4):895-908. doi:10.1016/j.cell.2012.12.041

20. Pattabiraman K, Golonzhka O, Lindtner S, et al. Article Transcriptional Regulation of Enhancers Active in Protodomains of the Developing Cerebral Cortex. *Neuron*. 2014;82(5):989-1003. doi:10.1016/j.neuron.2014.04.014

21. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10(12):1213-1218. doi:10.1038/nmeth.2688

22. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-)*. 2007;316(5830):1497-1502. doi:10.1126/science.1141319

23. Bonev B, Mendelson Cohen N, Szabo Q, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*. 2017;171(3):557-572.e24. doi:10.1016/j.cell.2017.09.043

24. Mo A, Mukamel EA, Davis FP, et al. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain NeuroResource Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*. 2015;86(6):1369-1384. doi:10.1016/j.neuron.2015.05.018

25. de la Torre-Ubieta L, Stein JL, Won H, et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell*. 2018;172(1-2):289-295.e18. doi:10.1016/j.cell.2017.12.014

26. Luo C, Lancaster MA, Castanon R, Nery JR, Knoblich JA, Ecker JR. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep*. 2016;17(12):3369-3384. doi:10.1016/j.celrep.2016.12.001

27. Nord AS, Blow MJ, Attanasio C, et al. Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell*. 2013;155(7):1521-1531. doi:10.1016/j.cell.2013.11.033

28. Amiri A, Coppola G, Scuderi S, et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science (80-)*. 2018;362(6420):eaat6720. doi:10.1126/SCIENCE.AAT6720

29. Markenscoff-Papadimitriou E, Whalen S, Przytycki P, et al. A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell*. 2020;0(0). doi:10.1016/j.cell.2020.06.002

30. Johnson MB, Kawasawa YI, Mason CE, et al. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*. 2009;62(4):494-509.

31. Nowakowski TJ, Bhaduri A, Pollen AA, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-)*. 2017;358(6368):1318-1323. doi:10.1126/science.aap8809

32. Tasic B, Yao Z, Graybuck LT, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*. 2018;563(7729):72-78. doi:10.1038/s41586-018-0654-5

33. Thomsen ER, Mich JK, Yao Z, et al. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat Methods*. 2015;13(1):87-93. doi:10.1038/nmeth.3629

34. Pollen AA, Nowakowski TJ, Chen J, et al. Molecular Identity of Human Outer Radial Glia during Cortical Development. *Cell*. 2015;163(1):55-67.

doi:10.1016/j.cell.2015.09.004

35. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561):486-490. doi:10.1038/nature14590

36. Cusanovich DA, Daza R, Adey A, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-)*. 2015;348(6237):910-914. doi:10.1126/science.aab1601

37. Preissl S, Fang R, Huang H, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci*. 2018:1. doi:10.1038/s41593-018-0079-3

38. Zhu C, Yu M, Huang H, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol*. 2019;26(11):1063-1070. doi:10.1038/s41594-019-0323-x

39. Lake BB, Chen S, Sos BC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol*. 2018;36(1):70-80. doi:10.1038/nbt.4038

40. Pollard KS, Salama SR, King B, et al. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet*. 2006;2(10):1599-1611. doi:10.1371/journal.pgen.0020168

41. Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc B Biol Sci*. 2013;368(1632):20130025-20130025. doi:10.1098/rstb.2013.0025

42. Reilly SK, Yin J, Ayoub AE, et al. Evolutionary changes in promoter and enhancer

activity during human corticogenesis. *Science (80-)*. 2015;347(6226):1155-1159. doi:10.1126/science.1260943

43. Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421-427. doi:10.1038/nature13595

44. Turner TN, Coe BP, Dickel DE, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*. 2017;171(3):710-722.e12. doi:10.1016/j.cell.2017.08.047

45. Stessman HAF, Xiong B, Coe BP, et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat Genet*. 2017;49(4):515-526. doi:10.1038/ng.3792

46. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*. 2020;180(3):568-584.e23. doi:10.1016/j.cell.2019.12.036

47. Li M, Santpere G, Imamura Kawasawa Y, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science (80-)*. 2018;362(6420):eaat7615. doi:10.1126/science.aat7615

48. Corces MR, Shcherbina A, Kundu S, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet*. 2020;52(11):1158-1168. doi:10.1038/s41588-020-00721-x

49. Lancaster MA, Renner M, Martin C-A, et al. Cerebral organoids model human brain development and microcephaly. *Nature*. 2013;501(7467):373-379.

doi:10.1038/nature12517

50. Pollen AA, Bhaduri A, Andrews MG, et al. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell*. 2019;176(4):743-756.e17. doi:10.1016/j.cell.2019.01.017

51. Camp JG, Badsha F, Florio M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci*. 2015;112(51):15672 LP - 15677. doi:10.1073/pnas.1520760112

52. Trevino AE, Sinnott-Armstrong N, Andersen J, et al. Chromatin accessibility dynamics in a model of human forebrain development. *Science (80-)*. 2020;367(6476):eaay1645. doi:10.1126/science.aay1645

53. Bhaduri A, Andrews MG, Mancia Leon W, et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature*. 2020;578(7793):142-148. doi:10.1038/s41586-020-1962-0

CHAPTER 2

**At the time of submission of this dissertation, this article was in press at Nature.*

Single cell epigenomics reveals mechanisms of human cortical development

Ryan S. Ziffra[1,2,3,4,5], Chang N. Kim[1,2,3], Jayden M. Ross[1,2,3], Amy Wilfert[6], Tychele N. Turner[7], Maximilian Haeussler[8], Alex M. Casella [9,10], Pawel F. Przytycki[11], Kathleen C. Keough[12,13], David Shin[1,2,3], Derek Bogdanoff[1,2,3], Anat Kreimer[4,5,14], Katherine S. Pollard[11,12,15,16,17], Seth A. Ament[9,18], Evan E. Eichler[6,19], Nadav Ahituv[4,5], Tomasz J. Nowakowski[1,2,3,17,*]

Affiliations:

1 Department of Anatomy, University of California, San Francisco, CA, USA

2 Department of Psychiatry, University of California, San Francisco, CA, USA

3 Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA, USA

4 Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

5 Institute for Human Genetics, University of California, San Francisco, CA, USA

6 Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

7 Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

8 Genomics Institute, University of California, Santa Cruz, CA, USA.

9 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

10 Medical Scientist Training Program, University of Maryland School of Medicine, Baltimore, MD

11 Gladstone Institutes, San Francisco, CA, USA

12 Institute for Computational Health Sciences, University of California, San Francisco, CA, USA

13 University of California, San Francisco, CA, USA

14 Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

15 Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

16 Quantitative Biology Institute, University of California, San Francisco, CA, USA

17 Chan Zuckerberg Biohub, San Francisco, CA, USA

18 Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD

19 Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

* Corresponding author: tomasz.nowakowski@ucsf.edu

SUMMARY

During mammalian development, chromatin state differences coincide with cellular differentiation and reflect changes in the gene regulatory landscape[1]. In the developing brain, cell fate specification and topographic identity play important roles in defining cell identity[2] and confer selective vulnerabilities to neurodevelopmental disorders[3]. To identify cell type specific chromatin accessibility patterns in the developing human brain, we used a single cell assay for transposase accessibility by sequencing (scATAC-seq) in primary human forebrain tissue samples. We applied unbiased analyses to identify genomic loci that undergo extensive cell type- and brain region-specific changes in accessibility during neurogenesis and an integrative analysis to predict cell type specific candidate regulatory elements. We found that cerebral organoids recapitulate most putative cell type-specific enhancer accessibility patterns but lack many cell type specific open chromatin regions found *in vivo*. Systematic comparison of chromatin accessibility across brain regions revealed an unexpected diversity among neural progenitor cells in the cerebral cortex and implicate retinoic acid signaling in the specification of prefrontal cortex neuronal lineage identity. Together, our results reveal the important contribution of chromatin state to the emerging patterns of cell type diversity and cell fate specification and provide a blueprint for evaluating the fidelity and robustness of cerebral organoids as a model for cortical development.

BACKGROUND

Cell types of the cerebral cortex (Figure 2.1a) have traditionally been classified based on a handful of morphological, anatomical, and physiological features. Recent innovations in single cell transcriptomics (scRNA-seq) have enabled massively parallel profiling of thousands of molecular features in individual cells and uncovered novel distinctions among closely related cell types, such as excitatory neurons located in different areas of the cerebral cortex[2,4]. Despite these advances, the developmental mechanisms underlying the emergence of distinct neuronal lineages in the human cerebral cortex remain largely unknown[5].

Chromatin state defines the functional architecture of the genome by modulating the accessibility of gene regulatory elements, such as enhancers, which serve as binding sites for transcriptional regulators. During development, sequential cascades of transcription factors progressively remodel and refine differential patterns of chromatin accessibility across distinct cell types[6,7]. Identifying the highly dynamic and cell type specific patterns of enhancer activity could provide critical insights into the molecular mechanisms that govern cell fate specification. Although chromatin accessibility represents a fundamental feature of cell identity, relatively few studies have profiled chromatin state changes during brain development[1,8–11]. Recently, the innovations in single cell genomics have enabled scalable profiling of chromatin state with cellular resolution using scATAC-seq[12]. In the developing mouse brain[13], scATAC-seq has revealed highly dynamic changed in chromatin accessibility underlying neurodevelopmental processes. Extending these studies to human primary tissue will be

needed to better understand how mutations in non-coding regulatory elements, including human-specific neurodevelopmental enhancers, interfere with normal developmental processes and contribute to genetic burden in psychiatric neurodevelopmental disorders[14,15].

RESULTS

Chromatin states of the developing brain

To characterize the chromatin state landscape of the developing human brain at single cell resolution, we performed scATAC-seq on primary samples of the human forebrain at mid-gestation (n = 6 individuals), including samples of dorsolateral prefrontal cortex (PFC), primary visual cortex (V1), primary motor cortex (M1), primary somatosensory cortex, dorso-lateral parietal cortex, temporal cortex, insular cortex, and the medial ganglionic eminence (MGE) (Figure 2.1b, Supplementary Table 1).

We generated data from 77,354 cells passing quality control criteria (Methods, Figure S2.1a-c). Aggregated signal from single cell libraries correlated strongly with bulk ATAC-seq libraries generated in parallel (Figure S2.1d), and biological replicates were highly correlated (Figure S2.1e-f). To reduce the dimensionality of the dataset, we performed latent semantic indexing followed by singular value decomposition (Methods). Batch correction was performed using the deep neural network-based scAlign[16] to correct for technical sources of variance, including individual variation and processing method (Figure S2.1g-k, Figure S2.2, Methods). We identified 25 distinct clusters using the Leiden

community detection algorithm (Figure 2.1c, Figure S2.1l-m, Methods). This analysis robustly separated cortical and subcortical cells (MGE) (Figure 2.1d).

To infer the identity of cell clusters, we calculated ‘gene activity scores’, which represents a proxy for gene expression[13], by summing fragments in the gene body and promoter regions (Methods). We identified the major cell classes, including radial glia (RG), intermediate progenitor cells (IPCs), deep layer (cortical layers V-VI) excitatory neurons (dlENs), upper layer (cortical layers II-IV) excitatory neurons (ulENs), MGE (IN-MGE) and CGE derived cortical interneurons (IN-CGE), insular neurons, progenitors from the MGE, microglia, oligodendrocyte progenitor cells (OPCs), endothelial cells, and mural cells (Figure 2.1e-f, Figure S2.3a). In addition, we used CellWalker[17] to assign cell type labels to scATAC-seq cells based on previously published scRNA-seq data (Methods, Figure S2.3b-c). CellWalker identified cell types at a finer resolution, including subtypes of broader cell classes. For example, radial glia form a single cluster, but multiple radial glia subtypes (‘dividing’, ‘ventricular’, ‘outer’, and ‘truncated’) are identifiable as sub-clusters using CellWalker (Figure S2.3d). Furthermore, we were able to identify differentially accessible peaks between two subclasses, tRGs and oRGs, that are distinguishable based on their expression of *CRYAB* and *HOPX*, respectively[2], suggesting that scATAC-seq is sensitive enough to distinguish cellular subtypes at high resolution (Figure S2.3e-f, Supplementary Tables 12-13,22).

Identifying cell type specific enhancers

To identify candidate gene regulatory elements, we called peaks on aggregate single cells from each broad cell class (Methods). Overlapping peaks were subsequently merged to a total union set of 459,953 peaks (Supplementary Tables 2-3). Annotation of our peak set in genomic features shows enrichment in intronic and distal intergenic regions and in the flanking regions of transcription start sites, suggesting an enrichment of gene regulatory elements, such as enhancers (Figure S2.4a-b). We intersected our peak set with the imputed 25-state chromatin model from Roadmap Epigenomics[18], finding strong enrichment for promoter and enhancer states and a depletion of transcribed, heterochromatin, and quiescent states (Methods, Figure S2.4c, Supplementary Table 14). Cell type specific differentially accessible peaks were identified for each cell type, resulting in a set of 265,123 peaks, with most cell types having on the order of thousands of specific peaks (Fisher's Exact, FDR<0.05, Figure 1g, Figure S2.4d-e, Supplementary Tables 4,6-7). In addition, we identified peaks that are differentially accessible between the eight brain regions used in this study (Figure S2.3g, Supplementary Table 8) In order to identify putative enhancers in our dataset, we integrated our ATAC-seq peaks with H3K27ac CUT&Tag data generated from similar samples (Methods), Hi-C data generated from developing human cortex[19], and gene expression data[2], using the Activity-by-contact algorithm[20] to predict enhancer-gene interactions (Methods) for all cortical cell types. In total, we predicted 25,659 gene-linked enhancers across the whole dataset (Figure S2.4d,f, Supplementary Table 5). We intersected our peaks with promoter-interacting regions identified using H3K4me3 PLAC-seq on sorted cells from developing human cortex[21], and found 67,493 peaks and 10,050 predicted enhancers with physical evidence

of promoter interaction (Figure S2.4d, Supplementary Table 15). Genes linked to predicted cell type specific enhancers were enriched for biological processes strongly associated with cell type identity (Methods, Figure S2.4g-h).

To further support our annotations, we intersected our peak set with publicly available datasets generated from human cortical tissue samples[9,22,23] (Figure S2.5a-c, Supplementary Table 15). We found that scATAC-seq recovered most of the peaks annotated using bulk tissue datasets, and also recovered many putative cell type specific peaks that are not captured in bulk datasets, especially those enriched in rarer cell populations, such as microglia and endothelial cells (Supplementary Table 4). We intersected our predicted enhancers with other enhancer predictions derived from previously published datasets[11,23,24]. Surprisingly, we did not find strong concordance between predicted enhancers from these studies (Figure S2.5d). Among functionally validated forebrain enhancers[25], the vast majority (304/319) overlapped chromatin accessibility peaks, but only 67 overlapped enhancers predicted using activity-by-contact (Figure 2.1h, Figure S2.5e). Together, these analyses suggest that scATAC-seq is a robust method for detecting chromatin accessibility patterns from heterogeneous tissue samples. However, limited overlap of predicted enhancers with previously published studies indicates that a better understanding of the relevant feature set to computationally predicting regulatory potential is urgently needed.

To characterize the regulatory ‘grammar’ of cell types, we calculated enrichment of known transcription factor binding motifs in cell type specific peak sets (Methods, Figure 2.1g,

Supplementary Table 20). We observed a strong association between transcription factor motif enrichments and cell type annotations from marker gene body enrichments. To examine transcription factor motif enrichments at the single cell level, we used ChromVAR[26] (Methods) and found substantial agreement with top motif enrichments for each cluster (Figure 2.4i). Together, these findings ascertain that scATAC-seq identifies chromatin accessibility patterns consistent with known transcription factor expression patterns across cell types and provides a roadmap towards discovery of transcription factor 'code' underlying cell lineage and cell fate specification.

Disease risk in the regulatory landscape

Mutations in non-coding genomic regions, as well as *de novo* loss of function mutations in chromatin regulators have been implicated in a wide range of neurodevelopmental and psychiatric disorders, including schizophrenia[27] and autism spectrum disorder (ASD)[3,28,29]. Cellular-resolution datasets of chromatin state across developmental stages and differentiation states may provide an important link between these mutations and selective vulnerabilities among the diverse cell types of the developing human brain, as was seen with recent studies using single cell transcriptomic data[3,22]. Towards that end, we intersected cell type specific ATAC-seq peaks, and putative enhancers, with disease-linked common and rare non-coding variants (Methods). We first intersected cell type specific peak sets, predicted enhancers, and peaks overlapping promoter-interacting regions[21] with genomic regions enriched for copy number variants in cases with developmental delay[30], identifying enrichment in dlEN, endothelial/mural, and microglia specific peaks, as well as peaks overlapping promoter-interacting regions in interneurons

(Figure 2.1i, Figure S2.6f). Because such regions do not provide specificity with respect to individual regulatory elements or genes, we next tested for enrichment of cell type specific peaks, predicted enhancers, and peaks overlapping promoter-interacting regions in the flanking regions of genes associated with ASD and NDD and identified peak sets significantly enriched and depleted in these regions for most cell types (Figure 2.1j, Figure S2.6a-c). We also intersected our cell type-specific peak sets and predicted enhancers with *de novo* non-coding mutations (DNMs) identified from ASD and neurodevelopmental delay (NDD) cases, however, no peak sets were significantly enriched for the currently annotated DNMs in probands compared to sibling controls (Fig S2.6d-e). In addition, we intersected predicted enhancers with topological associated domains that contain neurodevelopmental disease associated genes[24,31], identifying significant colocalization in TADs in several cell types (Figure S2.6g). Finally, we sought to assess the enrichment of common variants associated with neuropsychiatric disease risk in our predicted enhancers for each cell type. To do this we performed a partitioned heritability LD score regression analysis using summary statistics from large-scale genome-wide association studies of schizophrenia[27,32], ASD[33], major depressive disorder[34], and bipolar disorder[35] (Methods). We found that excitatory and inhibitory neuron putative enhancers were enriched (FDR < 0.05) for common variants associated with schizophrenia, confirming previous findings of neuronal involvement[27] (Figure 2.1k). Together, our prenatal cell type specific chromatin state data has the potential to identify specific regulatory programs during cortical development that confer the greatest risk for neurodevelopmental disorders, particularly as improved disease associated variant annotations become available.

Dynamic chromatin states of neurogenesis

To better understand how transcriptomic and epigenomic changes may regulate cell fate decisions during neurogenesis, we co-embedded scRNA-seq and scATAC-seq datasets for the relevant cell types generated from the visual cortex (Figure 2.2a-c, Methods). Projections of gene expression and gene activity scores in the co-embedded space reveals clustering of distinct cell types is preserved irrespective of the profiling modality (Figure 2.2d). To identify trajectories of chromatin accessibility underlying excitatory neuron differentiation and maturation, we performed pseudotemporal ordering of cells in the co-embedded space, which recovered the known developmental sequence of cell types underlying excitatory neuron differentiation (Figure 2.2e, Methods). We identified >25,000 peaks with transient accessibility across pseudotime, including >5,000 predicted enhancers, many of which are predicted to interact with genes linked to cell type identity (Figure 2.2f-h, Supplementary Table 9).

Consistent with recent reports[36,37], we find that, for genes with variable expression across pseudotime, gene activity scores derived from chromatin accessibility in the *cis*-regulatory region around genes are highly correlated with gene expression (Fig 2.2i, Figure S2.7, Methods). Finally, by calculating transcription factor binding site enrichment across peaks that show dynamic changes in accessibility along pseudotime, we reconstructed the known hierarchy of transcription factors involved in cortical neurogenesis, including sequential enrichment for PAX6, EOMES, and MEF2C binding sites among transiently accessible loci (Figure 2.2j). Together, these results underscore highly dynamic states of chromatin accessibility during human cortical neurogenesis.

Area specific chromatin states

Area-specific cortical excitatory neuron types emerge during early neurogenesis, but only limited transcriptomic differences have been found among progenitors from different regions[2]. Given that changes in the accessibility of regulatory elements often precede gene expression, we examined whether epigenomic signatures could foreshadow the emergence of area-specific excitatory neurons. Specifically, we compared scRNA-seq and scATAC-seq profiles of excitatory lineage cells sampled from the extremes of the rostral-caudal axis, PFC and V1 (Figure 2.3a-b, Figure S2.8a-h). For each modality, we ordered the cells in pseudotime to approximate the differentiation trajectory and identified the 'branch' point along this trajectory at which transcriptomic or chromatin state differences between PFC and V1 lineages become apparent (Figure 2.3c-f, Methods). In contrast to transcriptomic data, which has revealed area-specific clusters of excitatory neurons (Figure 2.3h), chromatin state signatures reveal a striking divergence between PFC and V1 intermediate progenitor populations (Figure 2.3g, Figure S2.8i-j). Transcriptomically, PFC and V1 IPCs differentially expressed only a handful of genes, including *NR2F1* (Supplementary Table 24), while chromatin accessibility analysis identifies over 1,800 differentially accessible peaks between these cell types (Figure S2.8k-l, Supplementary Table 25).

Next, to identify putative regulatory programs that could underlie the divergence of PFC and V1 lineages, we performed transcription factor binding site enrichment analysis on peaks that were differentially accessible between PFC and V1 cells (Fisher's Exact, two-sided, FDR<0.05, Figure 2.3i-k, Figure S2.9a-d, Supplementary Tables 10-11). This

analysis recovered several transcription factors predicted based on transcriptomic studies[2,38], including enrichment of POU3F2, MEIS1, TBR1, NEUROD1, NEUROG2, and TBX21 binding motifs among PFC cells. Interestingly, this analysis also identified components of the Retinoic Acid (RA) signaling pathway, including RXR, RAR and TGIF1, among PFC cells, consistent with a recent report that suggested increased levels of RA activity in the PFC during mid-gestation[39] (Figure 2.3j-k, Supplementary Table 21).

Retinoic acid in cortical arealization

RA signaling plays an important role in patterning of the neural tissue during mammalian brain development[39,40]. To test if RA may promote the differentiation of the human PFC lineages, we cultured cortical organoids in the presence, or absence, of Vitamin A, the precursor for RA synthesis and, in parallel, we treated a organoids that were cultured with Vitamin A with 4-diethylamniobenzaldehyde (DEAB), a potent inhibitor of RA synthesis[41] (Methods, Figure 2.3l). At week 10 of differentiation, which corresponds to deep layer neurogenesis, organoids were profiled using scRNA-seq. We found that excitatory forebrain neurons (*FOXG1/NEUROD2* double positive) cultured in the presence of Vitamin A clustered separately from those derived from organoids cultured without Vitamin A or in the presence of DEAB (Fig 2.3m-n, Figure S2.9e). Among the top differentially expressed genes, we found signatures that distinguish PFC and V1 cortical neurons, including *SATB2, NR2F1*, and *NR2F2*[2] (Figure 2.3n, Figure S2.9h-i). We applied a previously developed classifier for annotating PFC and V1 neuronal identities among organoid neurons[42], and found consistently higher proportion of neurons classified as PFC-like among organoids cultured with Vitamin A compared with those cultured without

Vitamin A or treated with DEAB (Figure S2.9f-g, Chi-square test, one-sided, *p-value < 0.00001). Differentially expressed genes, including co-expression among excitatory neurons of *SATB2/CTIP2* and *AUTS2* in prefrontal cortex, and enriched expression of NR2F1 in the visual cortex[2,43–45], were also confirmed by immunostaining and found to be consistent with a PFC-like identity of organoids cultured in the presence of Vitamin A (Figure 2.3n-o, Figure S2.9j). Together, these findings suggest a role for the retinoic signaling pathway in the specification of the PFC neuronal lineage during human cortical development, and further studies are required to determine how the RA pathway interfaces with other signaling pathways, such as fibroblast growth factor, to promote this neuronal fate[46].

Benchmarking cerebral organoids

Due to the scarcity of primary human tissue, studies of human neural development critically require suitable *in vitro* models, such as cerebral organoids. Previous studies emphasized the similarities between cerebral organoid cells and their *in vivo* counterparts using single cell transcriptomics[47,48] and bulk epigenomics[10,11,49]. We generated scATAC-seq data for 23,555 cells from cortical organoids derived via directed differentiation from three genetically normal individuals[47,50] at three time points of differentiation (Figure 2.4a-c, Figure S2.10a-e, Extended Data Table 1, Methods). To validate our organoid lines, we also generated scRNA-seq data from organoids derived from the same lines and cultured in parallel and showed that all lines expressed *FOXG1* and markers of major cell types (Figure S2.10j-n). Using gene activity scores, we identified the major cell classes of cell types among scATAC-seq data, including radial glia, IPCs, interneurons, and excitatory

neurons, although individual clusters contained fewer cell type specific peaks than clusters derived from primary cells (Figure 2.4d, Supplementary Table 17).

Next, we quantified chromatin accessibility among organoid cells across peaks defined from primary cells (Methods, Figure 2.4e) and found organoid cells homologous to radial glia, IPCs, interneurons, and excitatory neurons. Across organoid cells, we identified 377,448 peaks (Supplementary Table 16), and intersected them with the primary cell peaks, as well as a down-sampled set to match the cell abundance of organoid dataset (Methods, Figure 2.4f). Of the 459,953 peaks in the full primary dataset, 239,661 were also called in the organoid dataset (Figure 2.4g, Figure S2.10f, Supplementary Table 18). The set of peaks not detected in organoids was strongly enriched for peaks specific to cell types not found in significant numbers in our organoids, including microglia, endothelial cells, astrocytes, and OPCs (Figure S2.10h). After removing these peaks from analysis, the remaining peaks not found in organoids were enriched among V1 excitatory neurons. While cell type specific peaks identified in primary cells maintain cell type specificity in homologous organoid cell types, many of the cell type specific peaks were not detected (Figure 2.4h-i). However, the vast majority (>80%) of predicted enhancers were also identified in organoids, except for microglial enhancers (Fig 2.4i, Figure S2.10h). We also identified 109,960 organoid peaks not found in the primary cell dataset (Figure 2.4f, Figure S2.10g). Transcription factor motif enrichments analysis revealed that these organoid-specific peaks were enriched for HIF1A, HIF1B, and p53, which is consistent with the reported higher levels of cellular stress[42] (Supplementary Table 23). To further explore the robustness of our findings, we integrated our data with recently

published epigenomic datasets generated from organoids[11,49]. This integrative analysis revealed that 20,066 out of the 77,573 peaks found in organoids, but not primary tissue, were also detected in other published datasets (Figure S2.10f, Supplementary Table 19), suggesting that our findings can be independently validated.

DISCUSSION

In this study we profiled chromatin state of single cells of the developing human brain and found thousands of transiently accessible loci that track with neuronal differentiation. These states may reveal mechanisms governing the establishment of cell fate during neurogenesis, and intersecting them with comparable datasets from adult human brain may enable the complete reconstruction of the epigenomic neurodevelopmental trajectory [51]. Consistent with previous studies[52], intersection of chromatin state landscape with disease variants implicates post-mitotic, developing cortical excitatory neurons in the etiopathogenesis of neuropsychiatric disorders[9,22,53]. Future studies are needed to probe how disease-associated variants in these regulatory regions modify cell fate decisions in the developing cortex. By comparing the regulatory landscape of different cortical areas during development, we found distinct sets of transcription factor binding motifs differentially enriched between these two lineages. Our findings extend the well-established role of RA signaling in forebrain development[39,45] and suggest that RA signaling contributes to the specification of excitatory neurons of the human prefrontal cortex. Dysregulation of RA signaling has been implicated in a range of neurodevelopmental and psychiatric disorders[54], and therefore our findings may have implications for studies of these disorders.

METHODS

Tissue Source

De-identified tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations. Protocols were approved by the Human Gamete, Embryo, and Stem Cell Research Committee (institutional review board) at the University of California, San Francisco.

Nuclei isolation from fresh primary tissue

Cortical areas were microdissected from 3 specimens of mid-gestation human cortex, in addition to 3 specimens of non-area-specific mid-gestation human cortex. Tissue was dissociated in Papain containing Deoxyribonuclease I (DNase) for 30 minutes at 37C and samples were triturated to form a single cell suspension. 10^6 Cells were pelleted and lysed for 3 minutes in 100uL chilled Lysis Buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 0.1% Igepal CA-630, 0.01% Digitonin, 1% BSA). Lysed cells were then washed with 1mL chilled Wash Buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 1% BSA) and nuclei were pelleted at 500rcf for 5 minutes at 4C.

Nuclei isolation from frozen primary tissue

Tissue sections were snap frozen and stored at -80C. Nuclei were isolated from frozen tissues using the protocol published in Corces MR et al., 2017[55]. Briefly, frozen tissue samples were thawed in 2mL chilled Homogenization Buffer (10mM Tris pH7.8, 5mM $CaCl_2$, 3mM $Mg(Ac)_2$, 320 mM Sucrose, 0.1mM EDTA, 0.1% NP40, 167uM β-

mercaptoethanol, 16.7uM PMSF) and lysed in a pre-chilled dounce. Cell lysates were then centrifuged in an Iodixanol gradient for 20 minutes at 3000rcf at 4C in a swinging bucket centrifuge with the brake turned off. The nuclei band was then carefully pipetted and nuclei were diluted in Wash Buffer.

Cortical organoid differentiation protocol

Cortical organoids were cultured using a forebrain directed differentiation protocol[47,56]. Briefly, 2 genetically normal human induced PSC lines H28126 (Gilad Laboratory, University of Chicago) and 13234 (Conklin Laboratory, Gladstone Institutes), which were previously authenticated[47], and the embryonic stem cell line H1 (WiCell, authenticated at source), were expanded and dissociated to single cells using accutase. Cells tested negative for mycoplasma. After dissociation, cells were reconstituted in neural induction media at a density of 10,000 cells per well in 96 well v-bottom low adhesion plates. GMEM-based neural induction media includes 20% Knockout Serum Replacer (KSR), 1X non-essential amino acids, 0.11mg/mL Sodium Pyruvate, 1X Penicillin-Streptomycin, 0.1mM Beta Mercaptoethanol, 5uM SB431542 and 3uM IWR1-endo. Media was supplemented with 20uM Rock inhibitor Y-27632 for the first 6 days. After 18 days organoids were transferred from 96 to six well low adhesion plates and moved to an orbital shaker rotating at 90rpm and changed to DMEM/F12-based media containing 1X Glutamax, 1X N2, 1X B27 without Vitamin A and 1X anti-anti. At 35 days, organoids were moved into DMEM/F12-based media containing 1X N2, 1X B27 with Vitamin A and 1x anti-anti. Throughout culture duration organoids were fed every other day.

Nuclei isolation from cerebral organoids

Cerebral organoids were dissociated in Papain containing Deoxyribonuclease I (DNase) for 30 minutes at 37C and samples were triturated to form a single cell suspension. 10^6 Cells were pelleted and lysed for 3 minutes in 100uL chilled Lysis Buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 0.1% Igepal CA-630, 0.01% Digitonin, 1% BSA). Lysed cells were then washed with 1mL chilled Wash Buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 1% BSA) and nuclei were pelleted at 500rcf for 5 minutes at 4C.

Cortical Organoid Arealization Experiment

Two genetically normal iPSC lines (1323-4 & H28126) were differentiated into cortical organoids following the above protocol up to day 35. At day 35, organoids from each line were split into 3 different conditions: 1) normal media conditions for day 35 and beyond as described above (with Vitamin A), 2) normal media conditions for day 35 and beyond (with Vitamin A) plus 100uM 4-Diethylaminobenzaldehyde (DEAB), an inhibitor of retinoic acid synthesis, 3) normal media conditions for day 35 and beyond as described above except using B27 *without* Vitamin A. DEAB treatment was ended after one week, and culture conditions remain otherwise the same until day 70, at which time organoids were processed for scRNA-seq and fixed for immunohistochemistry. 13234 organoids were used for scRNA-seq (one for each of the three conditions) and both 13234 and H28126 organoids were used for immunostaining. Organoids processed for scRNA-seq were multiplexed using Multi-seq oligonucleotide barcoding[57] and pooled for library prep and sequencing to reduce potential batch effects.

Single Cell RNA-seq Library Preparation and Sequencing

Single cell RNA-seq libraries were generated using the 10x Genomics Chromium 3' Gene Expression Kit. Briefly, single cells were loaded onto chromium chips with a capture target of 10,000 cells per sample. Libraries were prepped following the provided protocol and sequenced on an Illumina NovaSeq with a targeted sequencing depth of 50,000 reads per cell. BCL files from sequencing were then used as inputs to the 10X Genomics Cell Ranger pipeline.

Bulk ATAC-seq Library Preparation and Sequencing

Bulk ATAC-seq libraries were generated using the protocol outlined in Corces MR et al., 2017[55]. Briefly, 50,000 nuclei were permeablized and tagmented. Tagmented chromatin libraries were generated and sequenced on an Illumina NovaSeq with a target sequencing depth of 50 million reads per library. Sequencing data was used as an input to the ENCODE ATAC-seq analysis pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline).

Bulk H3K27ac CUT&Tag Library Preparation and Sequencing

H3K27ac CUT&Tag libraries were prepared as previously described in Kaya-Okur et al., 2019[58], with modifications to the protocol. Briefly, cells were dissociated from human developing cortical tissue as described above. 50,000 cell aliquots were pelleted at 600xg in a swinging bucket rotor centrifuge and washed twice in 200µL CUT&Tag wash buffer (20 mM HEPES pH 7.5; 150 mM NaCl; 0.5 mM Spermidine; 1× Protease inhibitor cocktail (Roche)). Nuclei were isolated by resuspending cell pellets in 200µL Dig-wash buffer

(CUT&Tag wash buffer supplemented with 0.05% digitonin and 0.05% IGEPAL CA-630). Nuclei pellets were washed twice in 200µL Dig-wash buffer before resuspending in 100µL Dig-wash buffer supplemented with 2mM EDTA and a 1:50 dilution of H3K27ac primary antibody (Cell Signaling 8173) and incubated over night at 4°C on an overhead rotator. Excess primary antibody was removed by pelleting the nuclei at 600xg and washing twice in 200µL Dig-wash buffer. Secondary antibody (Novex A16031) was added at a dilution of 1:50 in 100µL Dig-wash buffer and nuclei were incubated at room temperature for 30 minutes rotating. Excess secondary antibody was removed by pelleting the nuclei at 600xg and washing twice in 200µL Dig-wash buffer. pA-Tn5 was added at a dilution of 1:100 in 100µL of Dig-med buffer (0.05% Digitonin, 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail), and nuclei were incubated at room temperature for 1 hour rotating. Unbound pA-Tn5 was removed by pelleting the nuclei at 300xg and washing twice in 200µL Dig-med buffer. Nuclei were resuspended 100µL tagmentation buffer (10 mM MgCl2 in Dig-med Buffer) and incubated for 1 hour at 37°C. After tagmentation, nuclei were lysed with the addition of 100µL DNA binding buffer (Zymo Research), and tagmented DNA was purified with a 1.5:1 ratio of AMPure XP beads (Beckman) following the manufacturer's instructions. Purified DNA was eluted in 21µL EB and mixed with 2µL each 10µM indexed i5 and i7 primers and 25µL NEBNext HiFi 2× PCR Master mix. Libraries were amplified with the cycling conditions: 72 °C for 5 min; 98 °C for 30 s; 12 cycles of 98 °C for 10 s and 63 °C for 30 s; final extension at 72 °C for 1 min and hold at 4 °C. Libraries were purified with a 1:1 ratio of AMPure XP beads and eluted in 15µL EB. CUT&Tag libraries were quantified by Agilent Bioanalyzer,

and sequenced paired-end to a depth of 15 million reads on a Illumina NovaSeq 6000 system, with read lengths 50x8x8x50.

Single Cell ATAC-seq Library Preparation and Sequencing

Nuclei were prepared as outlined in the 10X Genomics Chromium single cell ATAC-seq solution protocol (v1.0 kit was used). Nuclei were loaded with a capture target of 10,000 nuclei per sample. scATAC-seq libraries were prepared for sequencing following the 10X Genomics single cell ATAC-seq solution protocol. scATAC-seq libraries were sequenced using PE150 sequencing on an Illumina NovaSeq with a target depth of 25,000 reads per nucleus (Extended Data Table 1).

Single Cell ATAC-seq Analysis Pipeline

Cell Ranger

BCL files generated from sequencing were used as inputs to the 10X Genomics Cell Ranger ATAC pipeline. Briefly, FASTQ files were generated and aligned to GRCh38 using BWA. Fragment files were generated containing all unique properly paired and aligned fragments with MAPQ > 30. Each unique fragment is associated with a single cell barcode.

SnapATAC

Fragment files generated from the Cell Ranger ATAC pipeline were loaded into the SnapATAC[59] pipeline (https://github.com/r3fang/SnapATAC) and Snap files were generated. A cell-by-bin matrix was then generated for each sample by segmenting the

genome into 5-Kb windows and scoring each cell for reads in each window. Cells were filtered based on log(reads passed filters) between 3-5 and fraction of reads in promoters between 10-60% to obtain cells with high quality libraries. Bins were then filtered, removing bins overlapping ENCODE blacklist regions (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/). This matrix was then binarized and coverage of each bin was calculated and normalized by $\log_{10}$(count + 1). Z-scores were calculated from normalized bin coverages and bins with a z-score beyond ± 2 were filtered from further analysis. Cells with coverage of <500 bins were removed from the downstream analysis. A cell-by-cell similarity matrix was generated by calculating the Latent Semantic Index (LSI) of the binarized bin matrix. Singular value decomposition (SVD) was performed on the log TF-IDF matrix. The top 50 reduced dimensions were used for batch correction through scAlign.

scAlign Batch Correction

Multiple batches were integrated using the scAlign package[16] (https://github.com/quon-titative-biology/scAlign). The ATAC batches were first merged together to calculate the Latent Semantic Index (LSI) with the TF matrix log-scaled for input into SVD. The 50 reduced dimensions of LSI were used as inputs to the encoder. The latent dimension was set at 32 and ran with all-pairs alignment of all batches. The input dimension to the encoder was set to 50 to match the input dimensions and trained to 15,000 iterations using the small architecture setting with batch normalization (BN). The 32 dimensions were used for downstream analysis for finding neighbors. The scRNA-seq were processed using Seurat and computed the top 15 components from CCA for input into

scAlign, and the latent dimension was set to 20 using the small architecture with BN and 15,000 iterations. All alignments were unsupervised.

Clustering and Visualization

In order to visualize the high dimensionality dataset in 2D space, the latent dimensions for the ATAC and RNA data from scAlign were used to construct UMAP (https://arxiv.org/abs/1802.03426) graphs from Seurat. A K-nearest neighbor graph was constructed from the latent dimensions from scAlign using k=15. The leiden algorithm was then used to identify 'communities', or clusters, in the sample, representing groups of cells likely to be of the same cell type using resolution 0.8.

Calculating Gene Activity Scores

To create a proxy for gene expression, ATAC-seq fragments in the gene-body plus promoter (2Kb upstream from transcription start sites) of all protein-coding genes were summed for each cell to generate 'Gene Activity Scores'. A matrix was constructed for all gene activity scores by all cells. Due to the sparsity of scATAC-seq data, the MAGIC[60] imputation method was used, as implemented in the SnapATAC package, to impute gene activity scores based on the K-nearest neighbor graph.

Assigning Cell Type Labels to scATAC-seq Cells

Broad cell type classes were assigned to cells based on the gene activity scores of previously described cell type marker genes[2] (Figure S2.2a). To identify cell types at a higher resolutions, we assigned cell type labels to using the CellWalker[17] method, as

implemented in CellWalkR (v0.1.7). Briefly, we used CellWalker to integrate scRNA-seq derived labels from Nowakowski et al., 2017[2] with scATAC-seq data by building a network of label-to-cell and cell-to-cell edges and diffusing label information over this combined network to compensate for data sparsity in single cell data. We calculated cell-to-cell edge weight using the Jaccard similarity between cells. Label-to-cell weight was calculated as the sum of the products of the gene activity scores for that cell and the log fold change in expression level of each marker for that cell label. We tuned label edge weight using cell homogeneity as described in the CellWalker paper[17]. Diffusion resulted in a vector of influence scores of each label for each cell. We then smoothed these vectors for each cell by taking a weighted average of its scores with those of each of its ten closest neighbors (weighted such that each neighbor contributes one fifth as much as the cell in question) in UMAP space. Finally, we assigned cell type labels to each cell based on the label with the highest influence.

Peak Calling

Fragments from cells were grouped together by broad cell class (RG, IPC, ulEN, dlEN, EndoMural, AstroOligo, nEN, IN-MGE, IN-CGE, MGE progenitor, Insular, Microglia) and peaks were called on all cluster fragments using MACS2 (https://github.com/taoliu/MACS) with the parameters '--nomodel --shift -37 --ext 73 --qval 5e-2 -B --SPMR --call-summits'. Peaks from each cell type were then combined, merging overlapping peaks, to form a master peak set and a cell-by-peak matrix was constructed. This matrix was binarized for all downstream applications.

Determination of Differentially Accessible Peaks

Differentially accessible peaks for each cell type were determined by performing a two-sided Fisher's exact test and selecting peaks that had log fold change >0, and FDR-corrected p-value < 0.05, using the built in function in snapATAC 'findDAR'.

Visualizing Cluster Signal in Peaks

The deeptools suite[61] (https://deeptools.readthedocs.io/en/develop/) was used to visualize pileups of cluster-specific ATAC-seq signal (output from MACS2) in DA peak sets.

Intersection with 25 Chromatin State Model

To comprehensively categorize our peaks in genomic features genome-wide, we intersected our peak set with the 25-state model from the Roadmap Epigenomics Project[18], specifically using the data generated from sample E081, which was a sample of developing human brain. Enrichment of peaks within annotated regions of the genome was calculated using the ratio between the (#bases in state AND overlap feature)/(#bases in genome) and the [(#bases overlap feature)/(#bases in genome) X (#bases in state)/(#bases in genome)] as described previously[18].

Intersection with Epigenomic Datasets

We intersected our peak sets with several epigenomic datasets including ATAC-seq peaks from de la Torre-Ubieta et al., 2018[9] (GEO: GSE95023), ATAC-seq peaks from Markenscoff-Papadimitriou et al., 2020[23] (GEO: GSE149268), H3K4me3 PLAC-seq

promoter-interacting regions, generated from ENs, INs, IPCs, and RGs sorted from samples of developing human cortex[62], that were graciously provided by the author, H3K27ac peaks from Amiri et al., 2018[11] (taken from supplementary tables of publication), ATAC-seq peaks from Trevino et al., 2020[49] (GEO: GSE132403), H3K27ac peaks from Li et al., 2018[22] (obtained from http://development.psychencode.org), and high confidence enhancer predictions from Wang et al., 2018[24] (obtained from http://resource.psychencode.org/). Any peaks not already mapped to hg38 were lifted over using the UCSC LiftOver tool. Overlaps between peak sets were determined using the 'findOverlaps' function in R.

Transcription Factor Motif Enrichment Analysis

The findMotifsGenome.pl tool from the HOMER suite[63] (http://homer.ucsd.edu/homer/) was used to identify TF motif enrichments in peak sets. The ChromVAR[26] R package was used to identify TF motif enrichments at the single cell level in scATAC-seq data. Briefly, the peak-by-cell matrix from the snap object was used as an input, filtering for peaks open in at least 10 cells. Biased-corrected TF motif deviations were calculated for the set of 1,764 human TF motifs for each cell.

Enhancer-Gene Predicted Interactions

The Activity-by-Contact (ABC) model[20] (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) was used for prediction of enhancer-gene interactions from scATAC-seq data. Cell type specific ATAC-seq signal and peak outputs from MACS2 were used as inputs. Bulk H3K27ac CUT&Tag libraries generated from similar samples (see 'Bulk

H3K27ac CUT&Tag Library Preparation and Sequencing' section above) were used as a mark for active enhancers. Publicly available Hi-C data generated from similar samples[19] was used to demarcate regulatory neighborhoods, using the highest resolution available for each chromosome. Cell type specific gene expression profiles were generated from publicly available scRNA-seq data generated from similar samples[2] by averaging expression across each cell type. The default threshold of 0.02 was used for calling enhancer-gene interactions.

VISTA Enhancer Intersections

VISTA Enhancers were taken from the VISTA Enhancer Browser[25] (https://enhancer.lbl.gov/) and filtered for human sequences found to be active in the forebrain. Enhancers were lifted over to Hg38 using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and overlapping regions were merged, resulting in 319 unique regions. These regions were intersected with the peak set from all primary scATAC-seq cells and 304 peaks overlapping VISTA forebrain enhancer regions were identified.

Genomic Feature Annotations

The ChIPSeeker R package[64] (https://bioconductor.org/packages/release/bioc/html/ChIPseeker.html) was used to annotate all peak sets in genomic features.

Gene Ontology

Identification of enriched biological processes in the genes nearby sets of cell type specific enhancer predictions was performed using the GREAT alogrithm[65]. For each cell type, peaks that were both predicted enhancers and cell type specific were identified and enrichment of biological processes in the flanking genes of the regions relative to a background set of the full primary peak set was identified.

Calculating sample correlations

Correlation between samples was determined using the 'multiBamSummary' function from the deeptools python suite[61] on sample bam files. Bam file comparison was limited to the genomic space of the merged primary peak set (n=459,953 peaks), ignoring duplicates and unmapped reads. Heatmaps were then generated using the 'plotCorrelation' function.

scRNA-seq/scATAC-seq Coembedding

To anchor mRNA expression and chromatin state profiles in the same map of cell diversity, we applied scAlign on datasets where we profiled scRNA-seq and scATAC-seq in parallel in the same sample. This was achieved by linking gene expression data to gene activity scores derived from chromatin accessibility data. The gene activity scores were logRPM values derived from gene activity scores generated by the SnapATAC pipeline. Then the gene expression and gene activity scores were processed using Seurat, and then split into batches for input into scAlign. The encoder space was

computed using multi CCA of the 10 dimensions with latent dimensions at 20 using the 'small' architecture.

Pseudotime Analysis

The Monocle 3 R package[66] (https://cole-trapnell-lab.github.io/monocle3/) was used for pseudotime calculation of the co-embedded RNA and ATAC dataset. The radial glia cells were set as the root cells. The minimum branch length was 9 in the graph building. Monocle 3 was also used for the pseudotime calculation of the scRNA-seq PFC/V1 dataset. The Cicero package[67] (https://cole-trapnell-lab.github.io/cicero-release/) was used for the pseudotime calculation of the scATAC-seq PFC/V1 dataset.

Identification of Temporally Dynamic Peaks in the Excitatory Neuronal Lineage

scATAC-seq cells from V1 samples used in the co-embedding analysis were divided into ten equal bins by pseudotime. Average accessibility for each peak for each bin was determined. Peaks were considering temporally dynamic if they met the following criteria: accessible in a minimum of 10% of cells in the bin with the highest accessibility, accessible in a maximum of 20% of cells in the bin with the lowest accessibility, at least a difference of 10% in proportion of cells where the peak was accessible between the lowest and highest accessibility bins, and had an increase in proportion of accessibility in cells of at least 3x between the lowest and highest accessibility bins. In total 25,415/459,953 peaks met these criteria and were deemed to be temporally dynamic in the cortical excitatory neuronal lineage.

Comparison of Accessibility, Gene Expression, and TF Motif Enrichment Across Pseudotime

Since pseudotime was calculated on the co-embedded space of ATAC and RNA cells, we can directly compare temporal changes in gene expression, gene activity scores calculated from open chromatin, and transcription factor motif enrichment. For each of the genes, we calculated gene activity scores using Cicero[67] and calculated a 1,000 cell moving average across pseudotime from the ATAC cells. This value was normalized to represent a proportion of the maximum value. For gene expression, we calculated a 1,000 cell moving average across pseudotime from the RNA cells. This value was normalized to represent a proportion of the maximum value. For TF motif enrichment, using Z-scores from ChromVAR, we calculated a 1,000 cell moving average of the motif enrichment across pseudotime from the ATAC cells. LOESS regression lines were fit to the moving average data. For the generation of heatmaps, a similar approach was used, except values were averaged within 20 equally sized bins of pseudotime and normalized the maximum value.

Branchpoint Analysis

URD[68] (https://github.com/farrellja/URD/) was used to compare the branchpoint of ATAC and RNA independently. Deep-layer neurons weren't considered during this analysis due to obfuscating identities, and the batch corrected values were used as input to the diffusion map calculations to combat batch effects. Diffusion parameters were set to 150 nearest neighbors, and sigma was auto calculated from the data. The tree was

constructed using 200 cells per pseudotime bin, 6 bins per pseudotime window, and branch point p-value threshold of 0.001.

Identification of Homologous Cell Types in Primary and Organoid Samples

In order to identify homologous cell types between primary and organoid scATAC-seq datasets, reads from organoid cells were counted in peaks defined in the primary dataset, providing matching peak by cell matrices for primary and organoid datasets. DA peaks were then identified in each dataset for each cluster as described above, and the intersection of this DA peak set was used to calculate correlation between primary and organoid clusters after averaging peaks accessibility across all cells in each cluster. Homologous cell types were then determined based on the highest correlation values for each cluster.

Single Cell RNA-seq Analysis

Seurat

For primary samples used in Figs. 2-3, preprocessing of scRNA-seq data was conducted using a minimum of 500 genes and 5% mitochondrial cutoff was used and Scrublet[69] for doublet removal. The SCTransform[70] workflow in Seurat[71] were run separately on each batch. canonical component analysis (CCA) on the Pearson residuals from SCTransform was used as input into scAlign for batch correction. Dimensionality reduction and clustering were performed using PCA and leiden, respectively, using the default parameters of the Seurat pipeline. For organoid samples used in the arealization experiment in Figure 3, libraries from different conditions were demultiplexed using the

Multi-seq pipeline (https://github.com/chris-mcginnis-ucsf/MULTI-seq). The normal SCTransform workflow was then applied, as described above. Differentially expressed genes between conditions were identified using the 'FindMarkers' function with 'MAST' selected as the method. For organoid samples used for validation, scRNA-seq data were integrated following the Seurat SCTransform integration workflow using default parameters.

Classification of Area Identity of Organoid Cells

In order to systematically determine whether organoid cells had a transcriptomic identity more closely aligned with human PFC or V1 cells, we implemented a classifier method described in Bhaduri et al., 2020[42]. Briefly, area gene modules defined based on area-associated gene expression patterns[2,42] were generated and module eigengene values were determined for each organoid excitatory neuron using the 'moduleEigengenes' function from the WGCNA R package[72]. Organoids cells were then assigned an identity of 'PFC' or 'V1' based on the higher module eigengene value for each module. Significance of differences in proportions of identity labels between treatments were determined using a two-sided Chi-square test ($p < 0.05$).

Disease Intersection

De novo mutation (DNM) enrichment

Peak sets were intersected DNMs from 2,708 probands and 1,876 siblings using bedtools v2.24.0. DNMs were identified by an in-house pipeline. Briefly, variants from whole-genome sequencing data were called using four independent callers: GATK v3.8,

FreeBayes, Strelka, and Platypus. Variant calls from each caller were intersected, and filtered for read depth (> 9), allele balance (> 0.25), absence of reads supporting the mutation in parents, and identified by at least three of the four callers.

Sets of cell type specific peaks and peaks overlap PLAC-seq promoter-interacting regions were tested for an enrichment of DNMs in probands as compared to a background peak set which contained all primary peaks. We used a Fisher's exact test to compare the number of peaks with one or more DNMs between the cell type-specific peak set and the background peak set. We also performed a Wilcox rank sum test comparing the number of DNMs per peak in the cell type-specific set to the background peak set. We applied a Bonferroni multiple test correction to all p-values.

ASD/NDD gene set enrichment

We created gene plus upstream regulatory regions using bedtools v2.24.0, where we defined the upstream regulatory region as the 100Kb region upstream of the gene transcription start sites. Gene regions were defined using Gencode V27. The total number of peaks in each gene plus upstream regulatory region was quantified per gene for each cell type and compared to the number of peaks in the merged peak set for each gene set using a Fisher's exact test. The peaks in the remaining gene plus promoter regions were used as background. Gene sets from Coe et al., 2014[30] (COE253), Kaplanis et al., 2020[73] (DDD299) and SFARI gene (SFARI854) were used for enrichment testing. P-values were Bonferroni corrected for multiple tests (# of peak sets).

Morbidity map CNV enrichment

CNVs enriched in NDD cases from Coe et al., 2014 (n=70) were intersected with peak sets using bedtools 2.24.0; peaks were required to have a 50% overlap with the CNV region. The total number of peaks overlapping a CNV were compared to the number of peaks that did not overlap with a CNV for each cell type. The full primary peak set was used as background and compared by Fisher's exact test. P-values were Bonferroni corrected for multiple tests (# of peak sets).

Cell type-specific GWAS enrichment testing

We retrieved GWAS summary statistics for schizophrenia (Ripke et al., 2014[27]), bipolar disorder (Stahl et al., 2019[35]), and autism (Grove et al., 2019[33]) from the Psychiatric Genomics Consortium data portal (https://www.med.unc.edu/pgc). We also obtained GWAS summary statistics for schizophrenia (Pardiñas et al., 2018[32]) from http://walters.psycm.cf.ac.uk/. GWAS summary statistics for major depression (Howard et al., 2019[34]) were obtained from the authors under the auspices of a Data Use Agreement between 23AndMe and the University of Maryland Baltimore. We applied stratified LD score regression (LDSC version 1.0.1; Finucane et al., 2018[74]; Finucane et al., 2015[75]) to these summary statistics to evaluate the enrichment of trait heritability in each of 10 predicted enhancer sets. These associations were adjusted for the union of the peak sets as well as for 52 annotations from version 1.2 of the LDSC baseline model (including genic regions, enhancer regions and conserved regions; Finucane et al., 2015). Associations that met a cutoff of FDR < 0.05 were considered significant.

TAD Enrichment

Odds ratios were calculated as the likelihood of a TAD containing an ATAC peak if it also contained a gene from the set denoted by the subplot title, with significance identified using a Fisher's exact test. The magenta dotted line indicates a significance threshold of $p < 0.05$. Gene sets were obtained from http://resource.psychencode.org/[24,31]. TAD sets were from human brain, germinal zone (GZ) and cortical plate (CP) from Won et al., 2016[19].

Immunohistochemistry

Samples used for immunostaining were fixed in 4% PFA for 45 min, washed out with PBS, and incubated overnight in a 30% sucrose solution at 4c. Samples were then embedded in a 1:1 solution of OCT and 30% sucrose and frozen at -80c until ready for sectioning. Cryosections were prepared at a thickness of 16uM. Heat induced antigen retrieval was performed in 10mM sodium citrate (pH=6.0) for 15 min. Permeabilization was performed in PBS (pH=7.4) supplemented with 2% Triton X-100. Primary and secondary antibodies were diluted and incubated in PBS (pH=7.4) supplemented with 10% donkey serum, 2% Triton X-100, and 0.2% gelatin. Primary antibodies used in this study included: mouse anti-AUTS2 (1:200, Abcam ab243036), rabbit anti-NR2F1 (1:100, Novus Biologicals NBP1-31259), mouse anti-SATB2 (1:250, Santa Cruz Biotechnology SC-81376), rat anti-CTIP2 (1:500, abcam AB18465), rabbit anti-FOXG1 (1:500, Abcam ab196868), and rabbit anti-PAX6 (1:200, Biolegend 901301). Secondary antibodies used were AlexaFluor secondary antibodies. Images were collected using Leica SP5 confocal system and processed using ImageJ/Fiji.

AUTHOR CONTRIBUTIONS

R.S.Z. and T.J.N. designed the experiments. R.S.Z. and J.M.R. performed all experiments. D.B. generated CUT&TAG data. D.S. assisted in organoid differentiation. R.S.Z. and C.N.K. performed the majority of data analysis. P.F.P., K.C.K., and K.S.P. contributed to study design and performed data analysis. A.W., T.N.T., A.K., A.M.C., S.A.A., and E.E.E. performed the disease intersection analyses. R.S.Z., C.N.K., N.A, and T.J.N. wrote the paper with input from all authors. M.H. built the cell browser interface. T.J.N. and N.A. supervised the project.

COMPETING INTERESTS

We declare no competing financial interests related to this article.

ACKNOWLEDGEMENTS

We thank Aparna Bhaduri for help with generation of scRNA-seq data and its analysis, and helpful discussions throughout the project, as well as Madeline Andrews for sharing organoid cultures. We thank Michael Song and Yin Shen for graciously sharing PLAC-seq promoter-interacting regions for intersection with our data. We thank John Rubenstein for reading of the manuscript. This study was supported by NIH awards: psychENCODE award U01MH116438, Brain Initiative award U01MH114825, and the Psychiatric Cell Map Initiative Convergence Neuroscience award U01MH115747, NIMH award R01MH109907 (N.A. and K.S.P.), Autism Speaks Predoctoral Fellowship (11874 to R.S.Z.), Simons Foundation grant (SFARI 491371 to T.J.N.), NARSAD Young Investigator Grant (to T.J.N), gifts from Schmidt Futures and the William K. Bowes Jr. Foundation, the Gladstone Institutes, the National Institute of Mental Health (1K99MH117165 to T.N.T.), the US National Institutes of Health (NIH) (R01 MH101221 to E.E.E) and 5U41HG002371-19 to M.H.

DATA AVAILABILITY

scATAC-seq and scRNA-seq data derived from primary human samples are available on the NeMO archive (https://assets.nemoarchive.org/dat-gnot1gb) and the psychENCODE Knowledge Portal (https://www.synapse.org/#!Synapse:syn21392931). scATAC-seq and scRNA-seq data from derived from cortical organoids are also available on the psychENCODE Knowledge Portal (https://www.synapse.org/#!Synapse:syn21392931) and GEO (GSE163018). Peak level scATACseq primary and organoid ATAC-seq data is available through the UCSC Cell Browser (https://cortex-atac.cells.ucsc.edu/) and UCSC Genome Browser (https://urldefense.proofpoint.com/v2/url?u=https-3A__genome.ucsc.edu_s_Max_cortex-2Datac&d=DwIBaQ&c=iORugZls2LlYyCAZRB3XLg&r=wIGwA13tJ0H_yBH_8fGR_aHDv_Lb9BdBvaGRmKuMfC8&m=C-AKivMuKdU2JxBfFMIkS53e2NDAh9SJrG2tdmW5_MU&s=Sg2BoS6TTUoAMLyXiaM6hGHhNtG9LqaUBpXoPQxWBuQ&e=).

CODE AVAILABILITY

Custom codes used in this study are available at the following GitHub repository: https://github.com/NOW-Lab/scATACcortex.

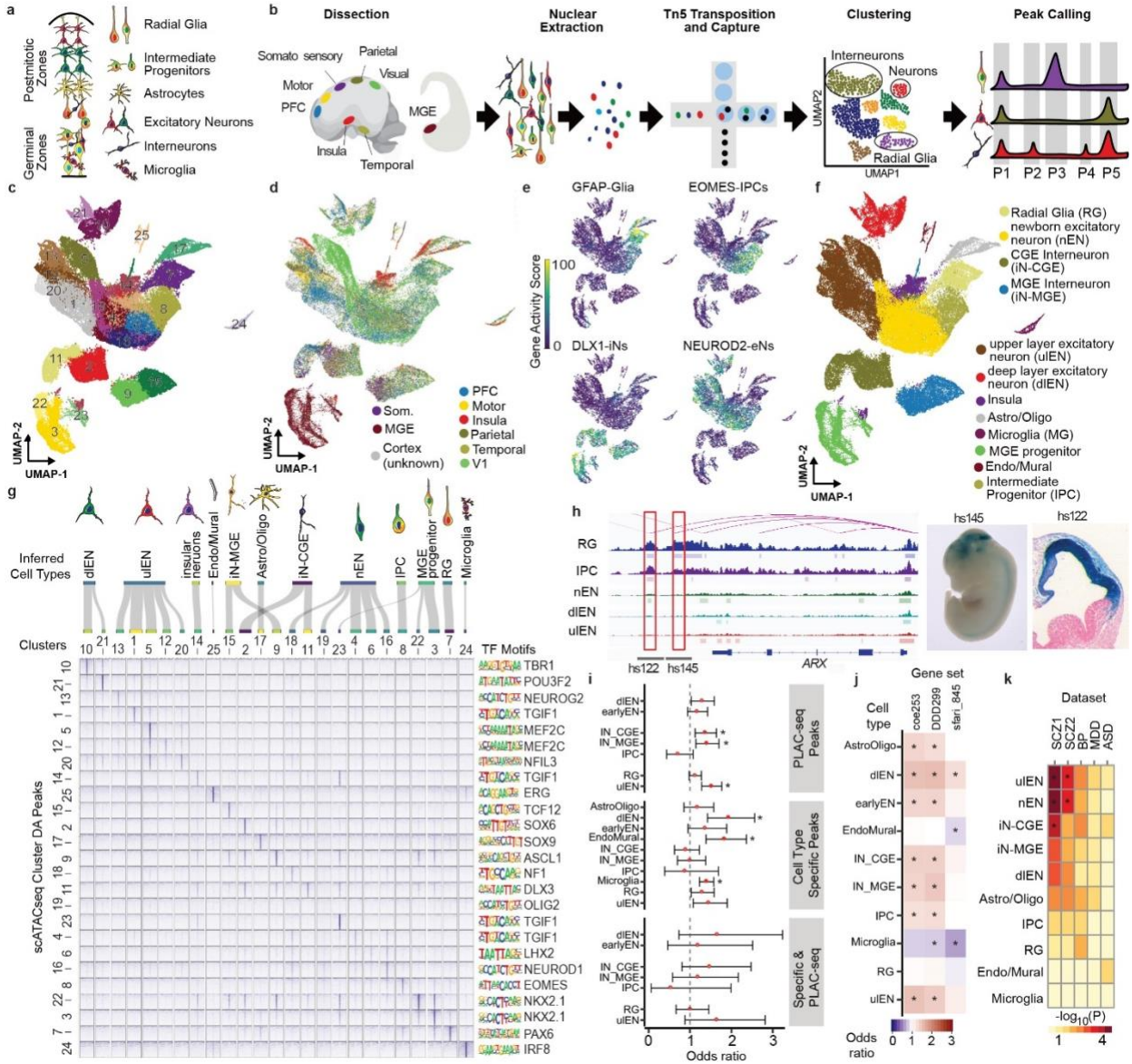

Figure 2.1: Single cell chromatin state atlas of the developing human brain. a) Schematic cross-section of developing cortex highlighting major cell types. **b)** Schematic depicting experimental workflow. PFC – prefrontal cortex, MGE – medial ganglionic eminence. **c)** UMAP projection of primary scATAC-seq cells (n = 6 individuals, 77,354 cells) colored by clusters. **d)** UMAP projection of primary scATAC-seq cells colored by brain region. Som. – somatosensory cortex, V1 – primary visual cortex. **e)** UMAP projections of gene activity scores for *GFAP* marking glia, *EOMES* marking intermediate progenitors, *DLX1* marking cells in the interneuron lineage, and *NEUROD2* marking cells in the excitatory neuron lineage. **f)** UMAP project of primary scATAC-seq cells colored by broad cell type. **g)** Top, sankey plot linking scATAC-seq clusters and cell type predictions. Bottom left, Pileups of ATAC-seq signal for each cluster within sets of the top 1000 enriched peaks for each cluster (Fisher's Exact, two-sided). Pileups are centered on peak centers and the +/-10Kb flanking region is depicted. Bottom right, significantly enriched TF motifs for each cluster specific peak set (Hypergeometric test, one-sided). **h)** Left,

predicted enhancer-gene interactions for radial glia highlighting predicted enhancers of ARX that overlap with validated VISTA forebrain enhancers[25]. Right, LacZ staining marking regions of enhancer activity for the enhancer candidates hs122 and hs145 in mouse embryos, depicting activity in the forebrain (Images taken from VISTA Enhancer Browser, https://enhancer.lbl.gov/) **i)** Enrichment and depletion of peaks that overlap with promoter interacting regions[21], cell type-specific peaks, and peaks that meet both criteria in copy number variant (CNV) regions enriched in pediatric cases of neurodevelopmental delay (NDD)[30] (n=70 NDD-associated CNVs) (Fisher's Exact, two-sided). Asterisks indicate Bonferroni corrected significance. Error bars represent 95% confidence interval. **j)** Enrichment and depletion of peaks that overlap with predicted enhancers in promoter and gene regions of genes associated for autism and NDD including genes enriched in *de novo* non-coding mutations (SFARI845, DDD299[73], COE253[30]) (Fisher's Exact, two-sided). Asterisks indicate tests that pass Bonferroni significance. **k)** Heatmap of heritability enrichment based on LD score regression analysis of GWAS summary statistics in cell type specific peak sets colored by $-\log_{10}(P)$. Asterisks indicate significance at FDR < 0.05. From left to right, Psychiatric Genomics Consortium (PGC) schizophrenia (SCZ) GWAS[27], an additional PGC schizophrenia GWAS[32], PGC bipolar (BIP) disorder[35], PGC major depressive disorder (MDD) GWAS[34], PGC autism spectrum disorder (ASD) GWAS[33].

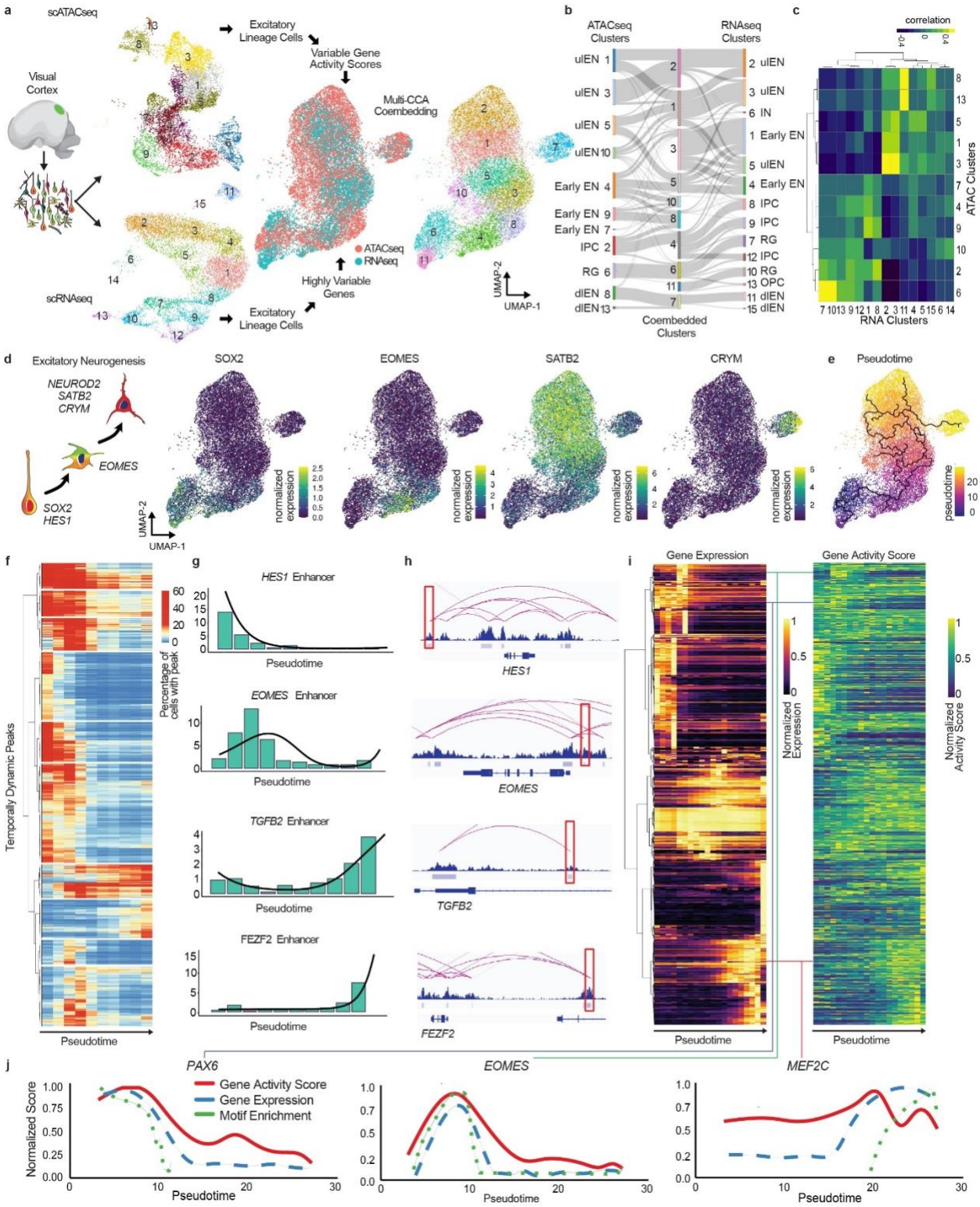

Figure 2.2: Dynamic changes in chromatin accessibility during human cortical neurogenesis. a) Schematic depicting workflow for co-embedding scATAC-seq and scRNA-seq data from the same samples. Left, schematic depicting experimental workflow. Top middle-left, UMAP projection of scATAC-seq cells from samples of visual cortex (n = 3 individuals) colored by leiden clusters. Bottom middle-left, UMAP projection

of scRNA-seq cell from samples of visual cortex (n = 2 individuals) colored by leiden clusters. Middle-right, UMAP projection of co-embedded cells colored by assay. Right, UMAP projection of co-embedded scATAC-seq & scRNA-seq cells colored by leiden clusters. **b)** Sankey plot depicting the mappings between scATAC-seq clusters, scRNA-seq clusters, and co-embedded clusters. **c)** Heatmap of correlations between scATAC-seq and scRNA-seq clusters based on a set of cell type marker genes. (Methods) **d)** Left, schematic depicting cell type marker genes in the cortical excitatory neuronal lineage. Right, projection of log normalized gene expression and gene activity scores in co-embedded space for *SOX2* (RGs), *EOMES* (IPCs), *SATB2* (Upper-layer ENs), and *CRYM* (Deep-layer ENs). **e)** UMAP projection of co-embedded cells colored by pseudotime with principal graph overlaid. **f)** Heatmap depicting the average proportion of cells with peaks that are differentially accessible across pseudotime (n=25,415). Cells are binned by pseudotime into 10 equally sized bins. **g)** Barplots of peak accessibility for 4 individual peaks across 10 pseudotime bins with regression line overlaid. **h)** Predicted enhancer-gene interactions for each of the four peaks, peaks highlighted in red. **i)** Heatmap depicting gene expression (left) and gene activity scores derived from open chromatin (right) for 615 cell type marker genes. Values are averaged within 20 equally sized bins of pseudotime. **j)** Comparison of moving averages of normalized gene activity scores (red), gene expression (blue), and motif enrichment (green) across pseudotime for *PAX6* (left), *EOMES* (middle), and *MEF2C* (right).

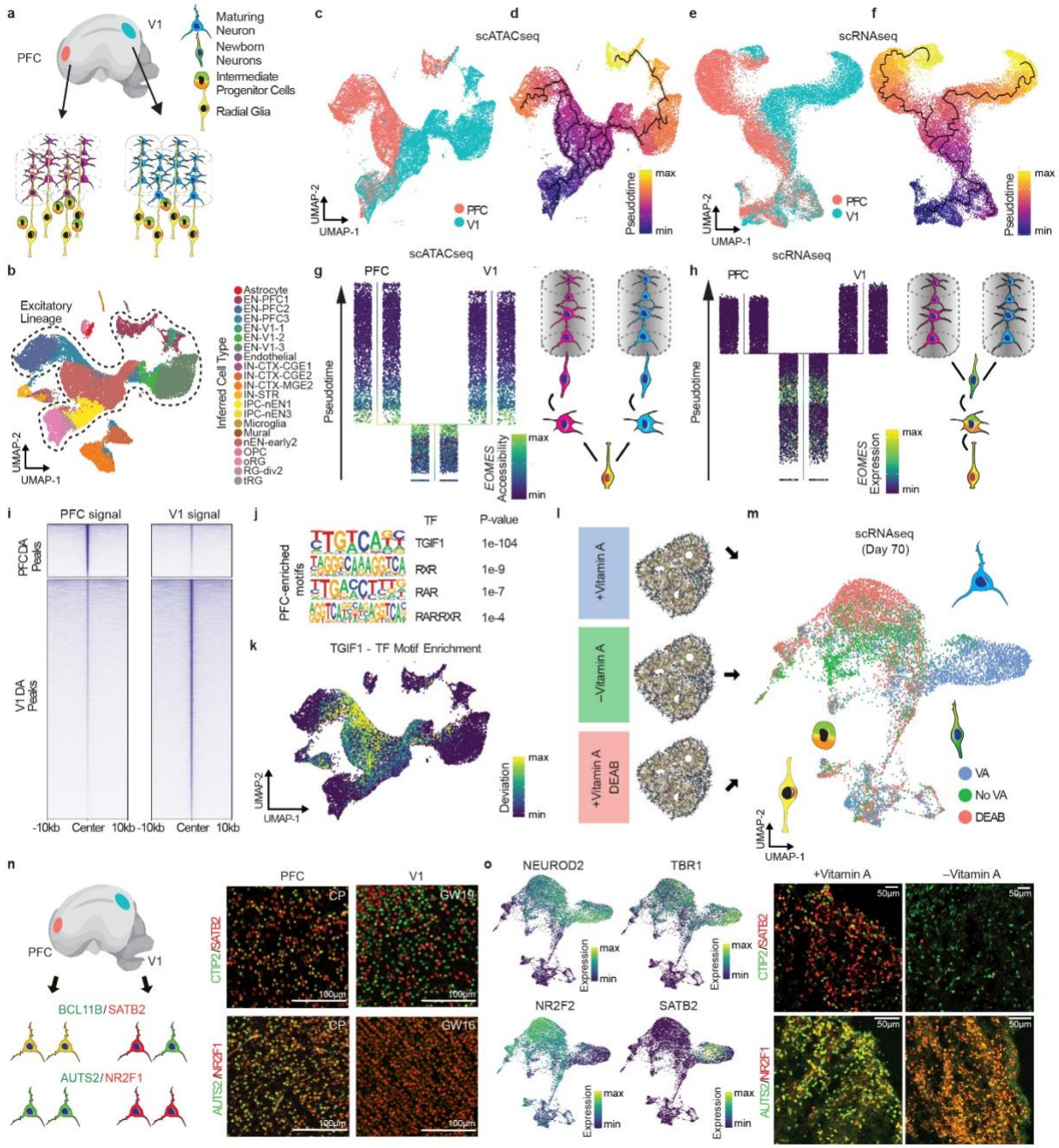

Figure 2.3: Areal differences in chromatin state of progenitor cells foreshadow the emergence of area-specific excitatory neuron types. a) Schematic depicting differentiation trajectories for excitatory neurons from the PFC (left) and V1 (right). **b)** UMAP projection of PFC & V1 scATAC-seq cells (n = 3 individuals) colored by cell type predictions. Cells from the excitatory lineage are outlined. **c)** UMAP projection of PFC & V1 scATAC-seq excitatory lineage cells colored by area of origin. **d)** UMAP projection of PFC & V1 scATAC-seq excitatory lineage cells colored by pseudotime value. **e)** UMAP projection of PFC & V1 scRNA-seq excitatory lineage cells (n = 2 individuals) colored by area of origin. **f)** UMAP projection of PFC & V1 scRNA-seq excitatory lineage cells colored

by pseudotime value. **g)** Left, PFC & V1 scATAC-seq excitatory lineage cells ordered from bottom to top by pseudotime value with PFC/V1 divergence branch point displayed (Methods). Cells colored by gene activity score of *EOMES*, highlighting IPCs. Right, schematic illustrating the excitatory neuron differentiation trajectory based on chromatin accessibility, in which PFC/V1 divergence becomes apparent at the level of IPCs. **h)** Left, Projection of PFC & V1 scRNA-seq excitatory lineage cells ordered from bottom to top by pseudotime value with PFC/V1 divergence branch point displayed (Methods). Cells colored by expression of *EOMES*, highlighting IPCs. Right, schematic illustrating the excitatory neuron differentiation trajectory based on gene expression, in which PFC/V1 divergence is not apparent in IPCs. **i)** Pileups of PFC and V1 signal in PFC and V1 specific peak sets. Pileups are centered on peaks showing +/-10Kb flanking regions. **j)** TF motif enrichments of retinoic acid-related TFs in set of 4,176 PFC specific peaks (Fisher's Exact, two-sided, FDR<0.05). **k)** UMAP projection of deviation scores of motif enrichment for TGIF1. **l)** Schematic of experimental design to test role of RA in organoid area identity. **m)** UMAP projection of scRNA-seq data from day 70 organoids (n=11,415 cells). Cells colored by treatment. **n)** Left, schematic of expected expression patterns of *BCL11B, SATB2, AUTS2*, and *NR2F1* in primary human cortex. Right, images of primary developing human cortex from the PFC (left) and V1 (right) stained for CTIP2/SATB2 (top) and AUTS2/NR2F1 (bottom). Representative images shown from n=2 specimens. **o)** Left, UMAP projection of cells colored by expression of *NEUROD2. TBR1, SATB2, NR2F2* (clockwise, from top left). Right, images of organoids cultured without Vitamin A (left) and with Vitamin A (right) stained for CTIP2/SATB2 (top) and AUTS2/NR2F1 (bottom). Representative images shown from n=3 lines.

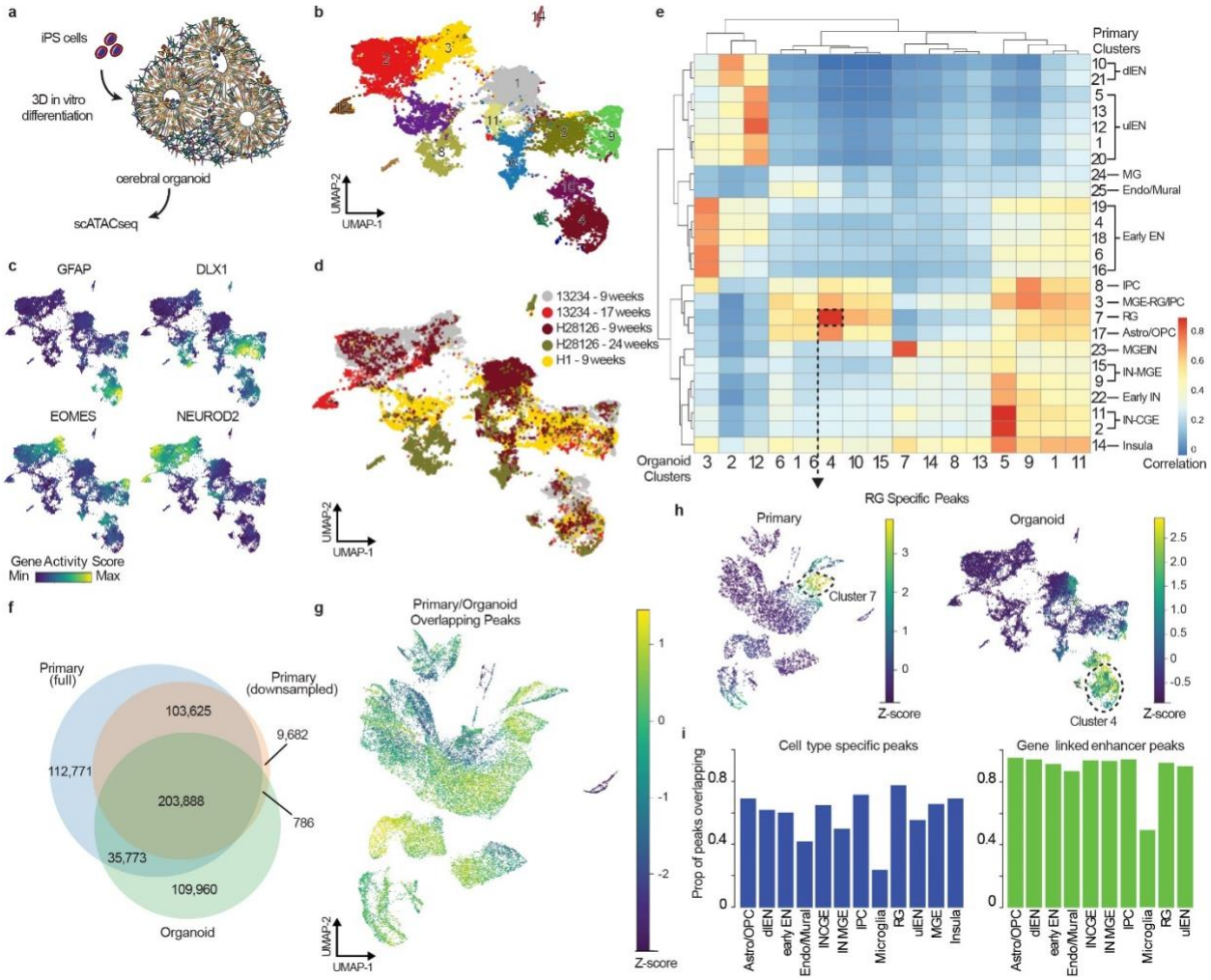

Figure 2.4: Cell type-specific differences in chromatin accessibility between cerebral organoids and the developing human brain. a) Schematic depicting experimental workflow. PSCs – pluripotent stem cells. **b)** UMAP projection of all organoid scATAC-seq cells (n = 5 organoids from 3 different lines and 3 different time points, 23,555 cells) colored by leiden clusters. (Cluster 16 not depicted, see Extended Data Figure 10e) **c)** UMAP projections of gene activity scores for *GFAP* marking radial glia, *EOMES* marking intermediate progenitors, *DLX1* marking interneurons, and *NEUROD2* marking excitatory neurons. **d)** UMAP projection of all organoid scATAC-seq cells colored by sample. **e)** Heatmap of pearson correlations between primary and organoid scATAC-seq clusters based on a common peak set. **f)** Venn diagram of overlap between the full primary peak set (red), a down-sampled primary peak set (blue), and the organoid peak set. **g)** UMAP project of enrichment Z-score of peaks that overlap between primary and organoid datasets on the primary scATAC-seq dataset **h)** Left, UMAP projection of enrichment Z-scores of radial glia specific peaks (Fisher's Exact, two-sided, FDR < 0.05) in all primary scATAC-seq cells. Right, UMAP projection of Z-scores of enrichment of the same radial glia specific peaks in all organoid scATAC-seq cells. **i)** Left, proportion of cell type specific primary peaks present in the organoid peak set. Right, proportion of gene-linked enhancers for each cell type present in the organoid peak set.

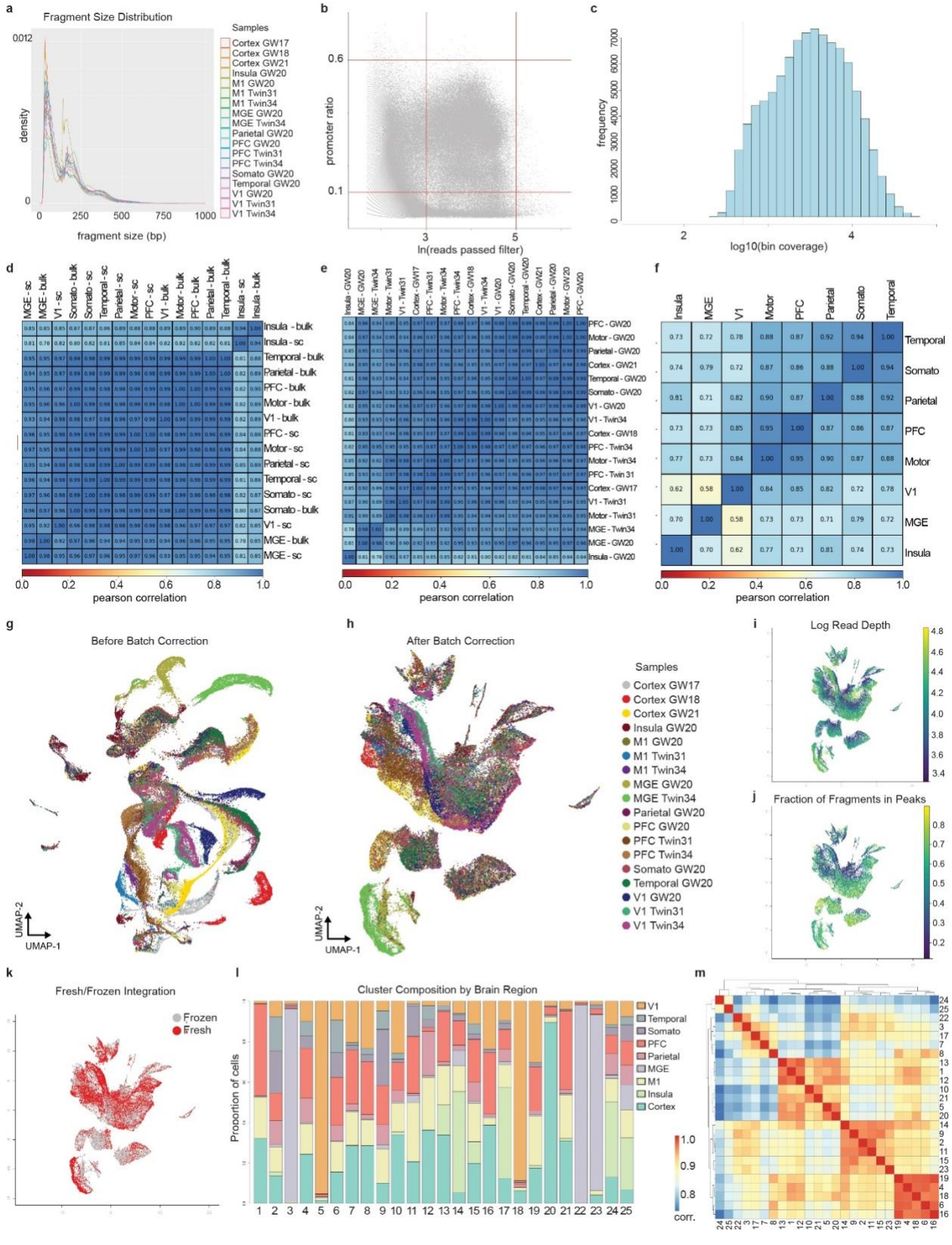
a
Fragment Size Distribution
density
fragment size (bp)
Samples
Cortex GW17
Cortex GW18
Cortex GW21
Insula GW20
M1 GW20
M1 Twin31
M1 Twin34
MGE GW20
MGE Twin34
Parietal GW20
PFC GW20
PFC Twin31
PFC Twin34
Somato GW20
Temporal GW20
V1 GW20
V1 Twin31
V1 Twin34
b
promoter ratio
ln(reads passed filter)
c
frequency
log10(bin coverage)
d
pearson correlation
e
pearson correlation
f
pearson correlation
g
Before Batch Correction
UMAP-1
UMAP-2
h
After Batch Correction
i
Log Read Depth
j
Fraction of Fragments in Peaks
k
Fresh/Frozen Integration
Frozen
Fresh
l
Cluster Composition by Brain Region
Proportion of cells
m
corr.

Figure S2.1: Batch correction and quality control metrics for primary scATAC-seq data. a) Density curve of fragment size distribution for deduplicated, uniquely mapped fragments passed quality filters for each sample. **b)** Plot of promoter ratio vs ln(read depth) for all cell barcodes detected. Cells were included in downstream analysis that had a promoter ration between 10-60% and ln(read depth) between 3-5. Red lines indicate upper and lower thresholds. **c)** Histogram of cellular coverage in 5kb genomic bins for all cells. Cells with reads in <500 bins were removed from downstream analysis **d)** Heatmap of all-by-all pearson correlations between all samples that had bulk libraries prepared in parallel. Bulk and aggregate single cell libraries for each sample are included. Correlation was calculated in the space of the merged primary peak set (n=459953 peaks). **e)** Heatmap of all-by-all pearson correlations between all aggregate single cell libraries. Correlation was calculated in the space of the merged primary peak set (n=459953 peaks). **f)** Heatmap of all-by-all correlations between primary single cell data aggregated by area of origin. Correlation was calculated in the space of the merged primary peak set (n=459953 peaks). **g)** UMAP projection of all primary scATAC-seq cells before batch correction colored by sample. **h)** UMAP projection of all primary scATAC-seq cells after batch correction colored by sample (Methods). **i)** UMAP projection of log(read depth) for deduplicated, uniquely mapped reads passed quality filters on all primary scATAC-seq cells **j)** UMAP project of fraction of fragments in peaks on all primary scATAC-seq cells. **k)** UMAP projection of all primary scATAC-seq cells colored by condition (fresh/frozen). **l)** Barplot depicting the proportion of cells from each brain region for each leiden cluster for all primary scATAC-seq cells. **m)** Heatmap of all-by-all pearson correlations between clusters. Correlation is calculated in the space of all primary peaks (n = 459953).

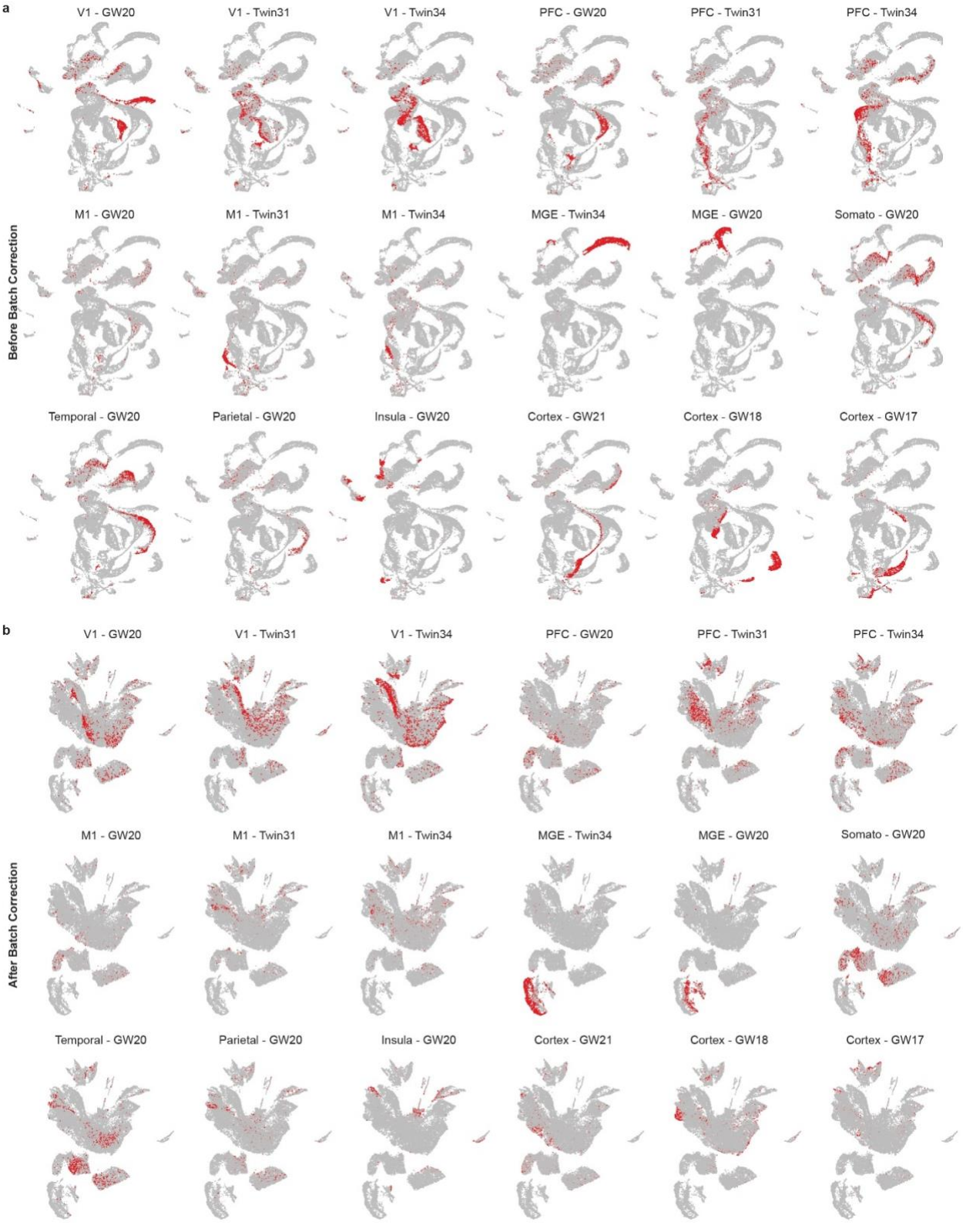
a
V1 - GW20
V1 - Twin31
V1 - Twin34
PFC - GW20
PFC - Twin31
PFC - Twin34
Before Batch Correction
M1 - GW20
M1 - Twin31
M1 - Twin34
MGE - Twin34
MGE - GW20
Somato - GW20
Temporal - GW20
Parietal - GW20
Insula - GW20
Cortex - GW21
Cortex - GW18
Cortex - GW17
b
V1 - GW20
V1 - Twin31
V1 - Twin34
PFC - GW20
PFC - Twin31
PFC - Twin34
After Batch Correction
M1 - GW20
M1 - Twin31
M1 - Twin34
MGE - Twin34
MGE - GW20
Somato - GW20
Temporal - GW20
Parietal - GW20
Insula - GW20
Cortex - GW21
Cortex - GW18
Cortex - GW17

Figure S2.2. Batch Correction of Primary scATAC-seq Samples. a) UMAP projection of all primary scATAC-seq cells passed QC before batch correction with all cells from each sample colored in red. **b)** UMAP projection of all primary scATAC-seq cells passed QC after batch correction with all cells from each sample colored in red.

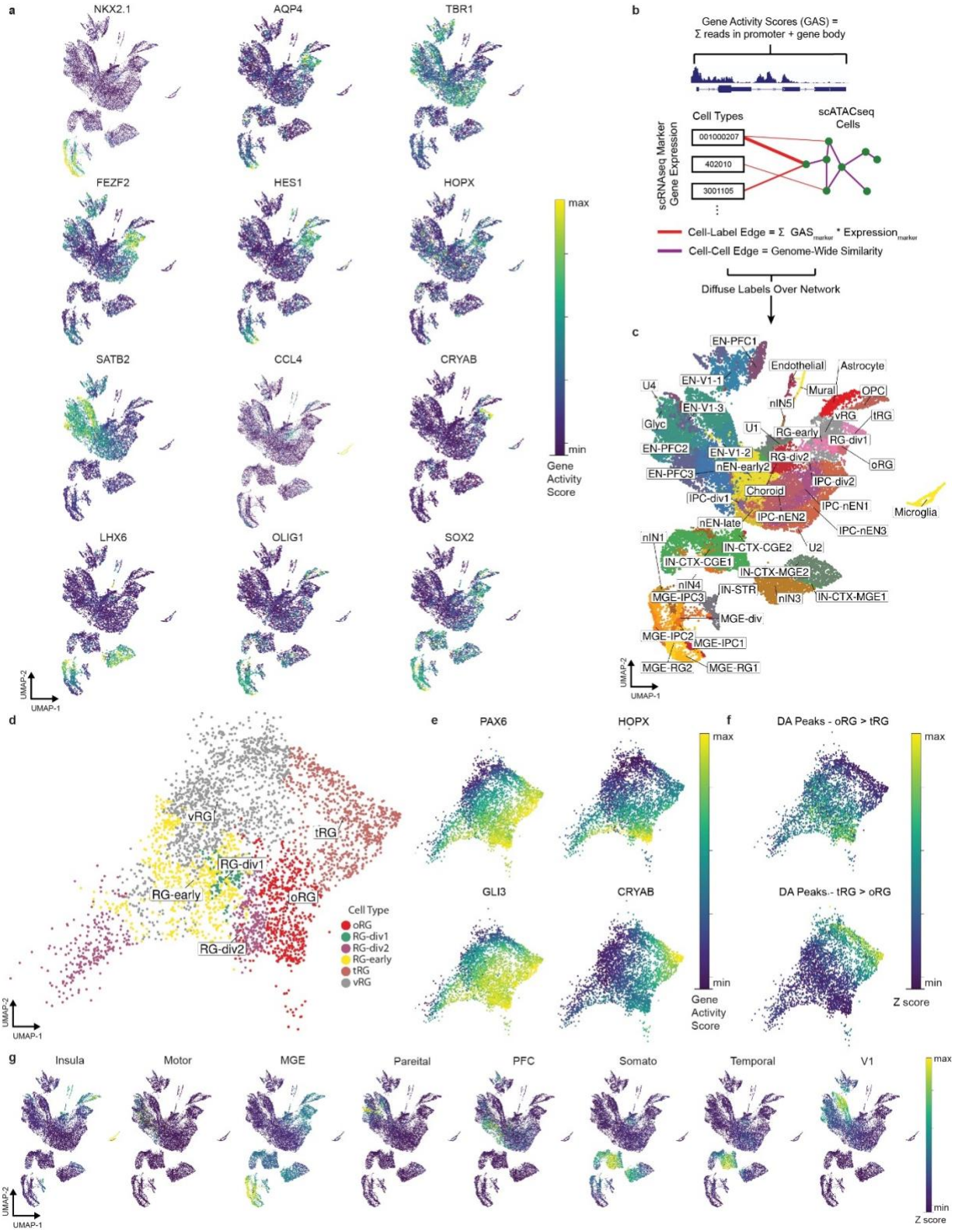
a
NKX2.1
AQP4
TBR1
FEZF2
HES1
HOPX
SATB2
CCL4
CRYAB
LHX6
OLIG1
SOX2
max
min
Gene Activity Score
UMAP-2
UMAP-1
b
Gene Activity Scores (GAS) =
Σ reads in promoter + gene body
Cell Types
scATACseq Cells
scRNAseq Marker Gene Expression
001000207
402010
3001105
Cell-Label Edge = Σ GAS marker * Expression marker
Cell-Cell Edge = Genome-Wide Similarity
Diffuse Labels Over Network
c
EN-PFC1
Endothelial
Astrocyte
U4
EN-V1-1
Mural
OPC
nIN5
EN-V1-3
vRG
tRG
Glyc
U1
RG-early
RG-div1
EN-PFC2
EN-V1-2
RG-div2
oRG
EN-PFC3
nEN-early2
IPC-div2
Choroid
IPC-div1
IPC-nEN1
Microglia
IPC-nEN2
nEN-late
IPC-nEN3
nIN1
IN-CTX-CGE2
U2
IN-CTX-CGE1
IN-CTX-MGE2
nIN4
IN-STR
nIN3
IN-CTX-MGE1
MGE-IPC3
MGE-div
MGE-IPC2
MGE-IPC1
MGE-RG2
MGE-RG1
UMAP-2
UMAP-1
d
vRG
tRG
RG-div1
RG-early
oRG
RG-div2
Cell Type
oRG
RG-div1
RG-div2
RG-early
tRG
vRG
UMAP-2
UMAP-1
e
PAX6
HOPX
GLI3
CRYAB
f
DA Peaks - oRG > tRG
DA Peaks - tRG > oRG
max
min
Gene Activity Score
Z score
g
Insula
Motor
MGE
Pareital
PFC
Somato
Temporal
V1
max
min
Z score
UMAP-2
UMAP-1

Figure S2.3: Gene activity scores correlate with cell type-specific expression of marker genes. a) UMAP projections of all primary scATAC-seq cells colored by gene activity score. From top left to bottom right, *NKX2.1* marking MGE cells*, AQP4* marking glia/astrocytes, *TBR1* marking excitatory neurons, *FEZF2* marking deep layer excitatory neurons, *HES1* marking radial glia*, HOPX* marking outer radial glia, *SATB2* marking upper layer excitatory neurons, *CCL4* marking microglia*, CRYAB* marking truncated radial glia, *LHX6* marking MGE-derived interneurons, *OLIG1* marking oligodendrocyte precursors, and *SOX2* marking radial glia. **b)** Schematic of the CellWalker algorithm[17] used to assign cell type labels to scATAC-seq cells based on integration with scRNA-seq data. **c)** UMAP projection of cell type labels assigned by CellWalker. **d)** UMAP projection of radial glia with cell type assignments from CellWalker. **e)** UMAP projections of gene activity scores for *PAX6* and *GLI3*, genes ubiquitously expressed in radial glia, *HOPX*, an oRG specific gene[2], and *CRYAB*, a tRG specific gene[2]. **f)** UMAP projection of Z scores of enrichment of oRG and tRG specific peaks (Fisher's Exact, two-sided test, P-value < 0.05). **g)** UMAP projections of Z-scores of enrichment of area specific peaks for each area in all primary scATAC-seq cells (Fisher's Exact, two-sided test, P-value < 0.05).

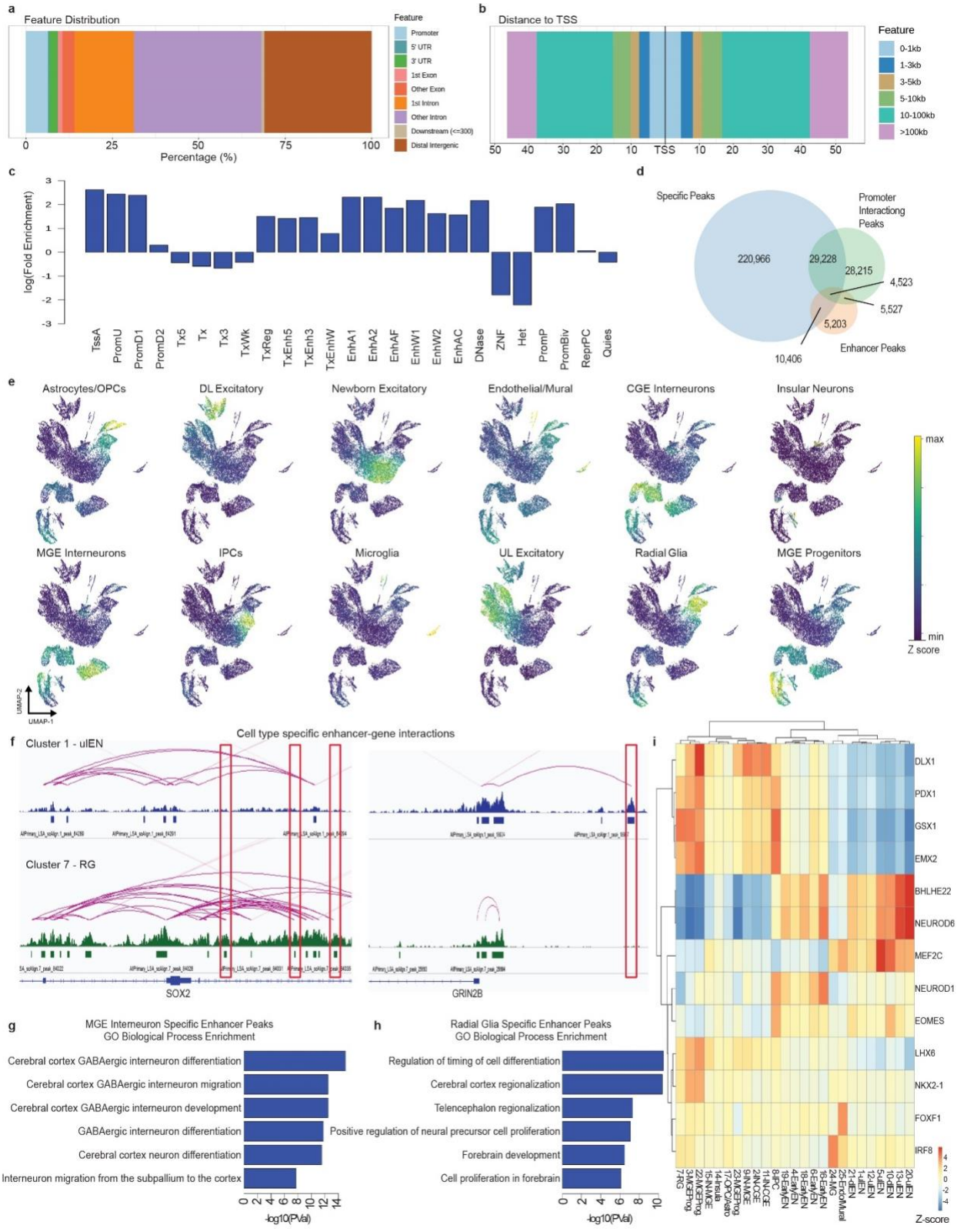
a
Feature Distribution
Feature
Promoter
5' UTR
3' UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
0
25
50
75
100
Percentage (%)
b
Distance to TSS
Feature
0-1kb
1-3kb
3-5kb
5-10kb
10-100kb
>100kb
50
40
30
20
10
TSS
10
20
30
40
50
c
log(Fold Enrichment)
3
2
1
0
-1
-2
-3
TssA
PromU
PromD1
PromD2
Tx5
Tx
Tx3
TxWk
TxReg
TxEnh5
TxEnh3
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAC
DNase
ZNF
Het
PromP
PromBiv
ReprPC
Quies
d
Specific Peaks
Promoter Interactiong Peaks
220,966
29,228
28,215
4,523
5,527
5,203
10,406
Enhancer Peaks
e
Astrocytes/OPCs
DL Excitatory
Newborn Excitatory
Endothelial/Mural
CGE Interneurons
Insular Neurons
MGE Interneurons
IPCs
Microglia
UL Excitatory
Radial Glia
MGE Progenitors
max
min
Z score
UMAP-2
UMAP-1
f
Cell type specific enhancer-gene interactions
Cluster 1 - uIEN
Cluster 7 - RG
SOX2
GRIN2B
g
MGE Interneuron Specific Enhancer Peaks
GO Biological Process Enrichment
Cerebral cortex GABAergic interneuron differentiation
Cerebral cortex GABAergic interneuron migration
Cerebral cortex GABAergic interneuron development
GABAergic interneuron differentiation
Cerebral cortex neuron differentiation
Interneuron migration from the subpallium to the cortex
0
2
4
6
8
10
12
14
-log10(PVal)
h
Radial Glia Specific Enhancer Peaks
GO Biological Process Enrichment
Regulation of timing of cell differentiation
Cerebral cortex regionalization
Telencephalon regionalization
Positive regulation of neural precursor cell proliferation
Forebrain development
Cell proliferation in forebrain
0
2
4
6
8
10
-log10(PVal)
i
DLX1
PDX1
GSX1
EMX2
BHLHE22
NEUROD6
MEF2C
NEUROD1
EOMES
LHX6
NKX2-1
FOXF1
IRF8
7-RG
22-MGEProg
3-MGEProg
15-MGE
14-Insula
17-OPC/Astro
23-MGEProg
9-NMGE
2-NCGE
11-NCGE
8-IPC/EN
19-EarlyEN
4-EarlyEN
18-EarlyEN
6-EarlyEN
16-EarlyEN
24-MG
25-Endo/Mural
21-dIEN
1-uIEN
12-dIEN
5-uIEN
10-dIEN
13-uIEN
20-uIEN
4
2
0
-2
-4
Z-score

Figure S2.4: Annotation of primary scATAC-seq peaks. a) Distribution of primary scATAC-seq peaks in genomic features. **b)** Distribution of primary scATAC-seq peaks are transcription start sites. **c)** Barplot of log(fold enrichment) of primary scATAC-seq peaks in chromatin states. Chromatin states defined by the 25-state model from Roadmap Epigenomics[18] (Methods). **d)** Intersection of cell type specific peaks (Fisher's Exact, two sided, FDR < 0.05), predicted enhancer peaks (Methods), and peaks overlapping promoter-interacting regions identified by H3K4me3 PLAC-seq[21]. **e)** UMAP projection of Z-scores of enrichment of cell type specific peaks (Fisher's Exact, two-sided, FDR < 0.05) for each broad cell type. **f)** Browser tracks highlighting cell type specific predicted enhancers. Left, highlighting predicted enhancers linked to *SOX2* in RGs that are not present in ulENs. Right, highlighting a predicted enhancer for *GRIN2B* that is present in ulENs and not RGs. **g)** Barplot of -log10(P) of Gene ontology biological processes that are enriched in cell type specific predicted enhancers of MGE-derived interneurons (Methods). **h)** Barplot of -log10(P) of Gene ontology biological processes that are enriched in cell type specific predicted enhancers of RGs (Methods). **i)** Heatmap of Z-scores of transcription factor motif enrichments of key lineage-associated TFs in each cluster.

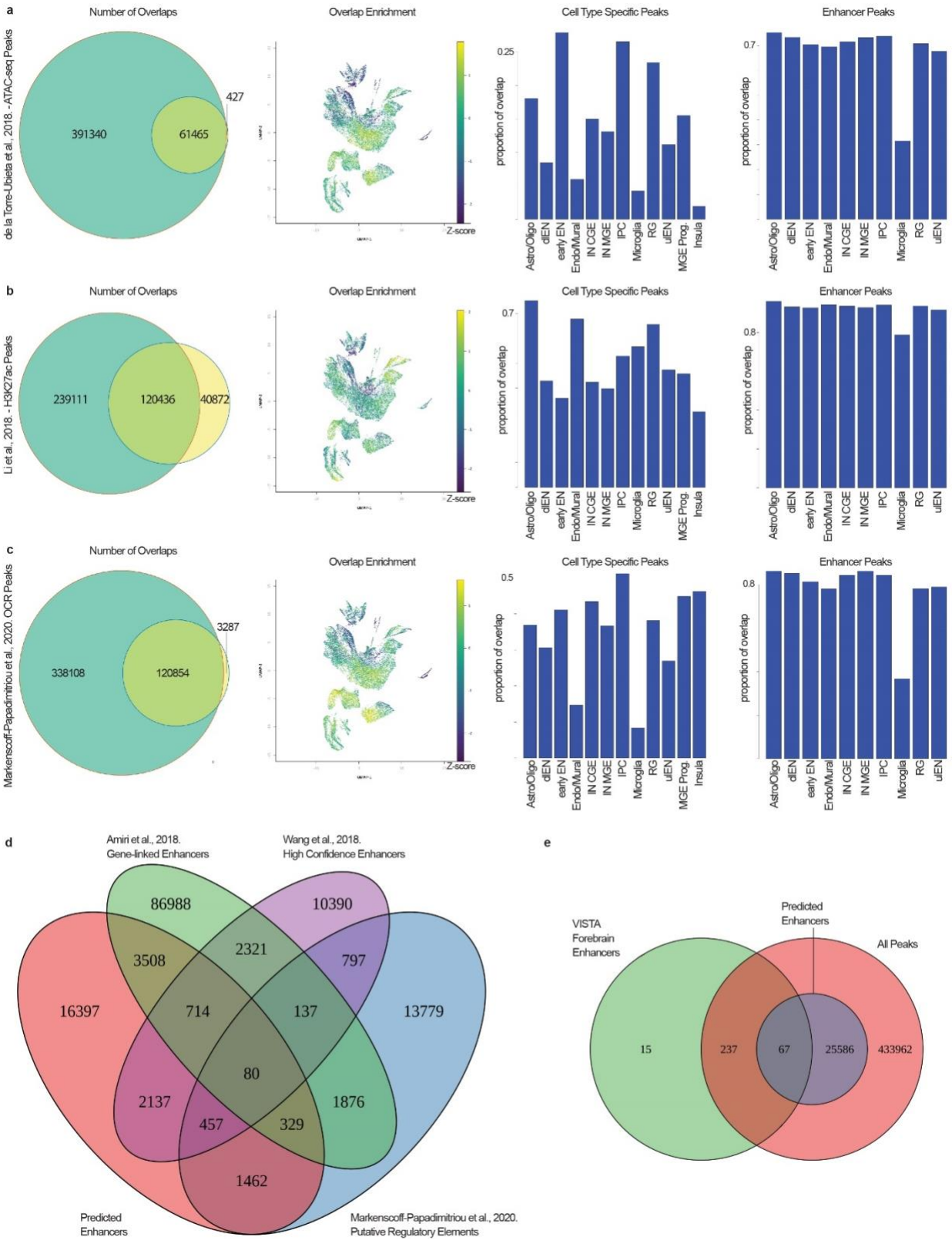
a
Number of Overlaps
Overlap Enrichment
Cell Type Specific Peaks
Enhancer Peaks
de la Torre-Ubieta et al., 2018. -ATAC-seq Peaks
391340
61465
427
Z-score
proportion of overlap
Astro/Oligo
dIEN
early EN
Endo/Mural
IN CGE
IN MGE
IPC
Microglia
RG
uIEN
MGE Prog.
Insula
b
Number of Overlaps
Overlap Enrichment
Cell Type Specific Peaks
Enhancer Peaks
Li et al., 2018. - H3K27ac Peaks
239111
120436
40872
Z-score
c
Number of Overlaps
Overlap Enrichment
Cell Type Specific Peaks
Enhancer Peaks
Markenscoff-Papadimitriou et al., 2020. OCR Peaks
338108
120854
3287
Z-score
d
Amiri et al., 2018.
Gene-linked Enhancers
Wang et al., 2018.
High Confidence Enhancers
86988
10390
3508
2321
797
16397
714
137
13779
80
2137
1876
457
329
1462
Predicted
Enhancers
Markenscoff-Papadimitriou et al., 2020.
Putative Regulatory Elements
e
VISTA
Forebrain
Enhancers
Predicted
Enhancers
All Peaks
15
237
67
25586
433962

Figure S2.5: scATAC-seq peaks overlap with previously annotated bulk ATAC-seq peaks and validated forebrain enhancers. a) Overlap of all primary peaks with the peak set from de la Torre-Ubieta et al., 2018[9]. (Left to right) Venn diagram of overlaps, UMAP projection of Z-scores of enrichment of overlapping peaks in all primary scATAC-seq cells, barplot of proportions of cell type specific peaks (Fisher's Exact, two-sided, FDR < 0.05) present in overlapping set, barplot of proportions of predicted enhancers for each cell type (Methods) present in the overlapping set. **b)** Overlap of all primary peaks with the peak set from Li et al., 2018[22]. (Left to right) Venn diagram of overlaps, UMAP projection of Z-scores of enrichment of overlapping peaks in all primary scATAC-seq cells, barplot of proportions of cell type specific peaks (Fisher's Exact, two-sided, FDR < 0.05) present in overlapping set, barplot of proportions of predicted enhancers for each cell type (Methods) present in the overlapping set. **c)** Overlap of all primary peaks with the peak set from Markenscoff-Papadimitriou et al., 2020[23]. (Left to right) Venn diagram of overlaps, UMAP projection of Z-scores of enrichment of overlapping peaks in all primary scATAC-seq cells, barplot of proportions of cell type specific peaks (Fisher's Exact, two-sided, FDR < 0.05) present in overlapping set, barplot of proportions of predicted enhancers for each cell type (Methods) present in the overlapping set. **d)** Venn diagram of overlap of all predicted enhancers (Methods) with gene-linked enhancers from Amiri et al., 2018[11], high confidence enhancers from Wang et al., 2018[24], and putative regulatory elements from Markenscoff-Papadimitriou et al., 2020[23]. **e)** Venn diagram of overlaps of VISTA forebrain enhancers[25] with all primary scATAC-seq peaks, and all predicted enhancers (Methods).

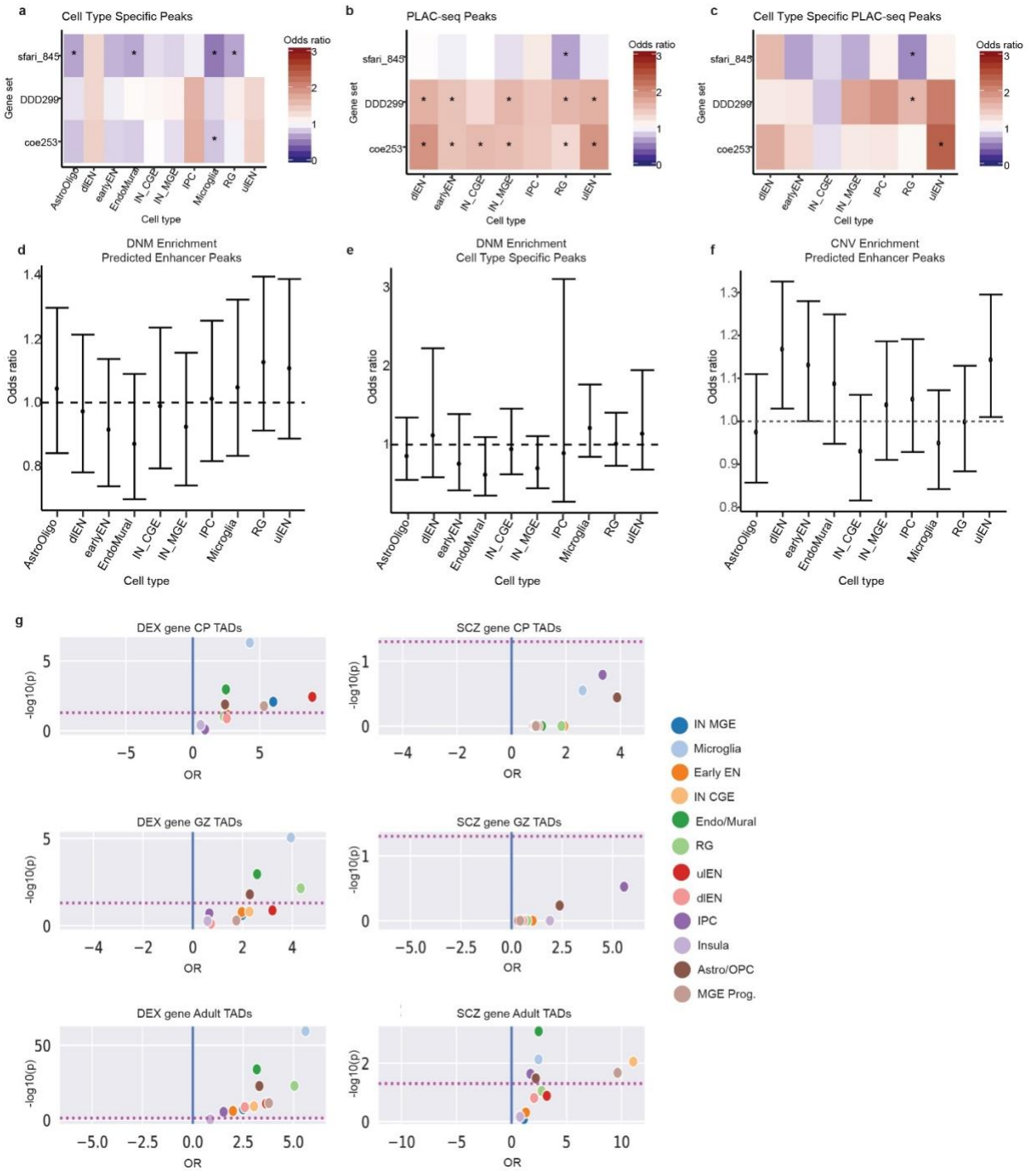
a
Cell Type Specific Peaks
b
PLAC-seq Peaks
c
Cell Type Specific PLAC-seq Peaks
Gene set
sfari_845
DDD299
coe253
Odds ratio
Cell type
AstroOligo
dIEN
earlyEN
EndoMural
IN_CGE
IN_MGE
IPC
Microglia
RG
ulEN
d
DNM Enrichment
Predicted Enhancer Peaks
e
DNM Enrichment
Cell Type Specific Peaks
f
CNV Enrichment
Predicted Enhancer Peaks
g
DEX gene CP TADs
SCZ gene CP TADs
DEX gene GZ TADs
SCZ gene GZ TADs
DEX gene Adult TADs
SCZ gene Adult TADs
-log10(p)
OR
IN MGE
Microglia
Early EN
IN CGE
Endo/Mural
RG
ulEN
dIEN
IPC
Insula
Astro/OPC
MGE Prog.

Figure S2.6. Enrichment and depletion of disease associated variants in scATAC-seq peaks. a) Enrichment and depletion of cell type specific peaks (Fisher's Exact, two-sided, FDR < 0.05) in promoter and gene regions of genes associated for autism and NDD including genes enriched in de novo non-coding mutations (DNM) (SFARI845, DDD29960[73], COE25361[30]). Stars indicate tests that pass Bonferroni significance. **b)** Enrichment and depletion of peaks that overlap with H3K4me3 PLAC -seq interactions[21] (Fisher's Exact, two-sided, FDR < 0.05) in promoter and gene regions of genes associated for autism and NDD including genes enriched in *de novo* non-coding mutations (DNM) (SFARI845, DDD29960[73], COE25361[30]). Stars indicate tests that pass Bonferroni significance. **c)** Enrichment and depletion of cell type specific peaks that overlap with H3K4me3 PLAC -seq interactions[21] (Fisher's Exact, two-sided, FDR < 0.05) in promoter and gene regions of genes associated for autism and NDD including genes enriched in *de novo* non-coding mutations (DNM) (SFARI845, DDD29960[73], COE25361[30]). Stars indicate tests that pass Bonferroni significance. **d)** Enrichment and depletion of DNMs in predicted enhancer peaks for each cell type in ASD probands compared with unaffected siblings. DNM data from a total of 2767 probands and 1855 unaffected siblings were included in the analysis. No tests reached Bonferroni significance. Bars represent 95% confidence interval. **e)** Enrichment and depletion of DNMs in cell type specific peaks for each cell type in ASD probands compared with unaffected siblings. DNM data from a total of 2767 probands and 1855 unaffected siblings were included in the analysis. No tests reached Bonferroni significance. Bars represent 95% confidence interval. **f)** Enrichment and depletion of predicted enhancer peaks in copy number variant (CNV) regions enriched in pediatric cases of neurodevelopmental delay (NDD)[30] (n=70 NDD-associated CNVs). No tests reached Bonferroni significance. Bars represent 95% confidence interval. **g)** Enrichment of cell type specific enhancers located in TADs with neurodevelopmental disease associated genes[31] (left) and schizophrenia associated genes[24] (right). Three distinct sets of TADs were used (top to bottom): TADs defined in the cortical plate (CP) of developing human cortex, TADs defined in the germinal zone (GZ) of developing human cortex, and TADs defined in adult cortex[19].

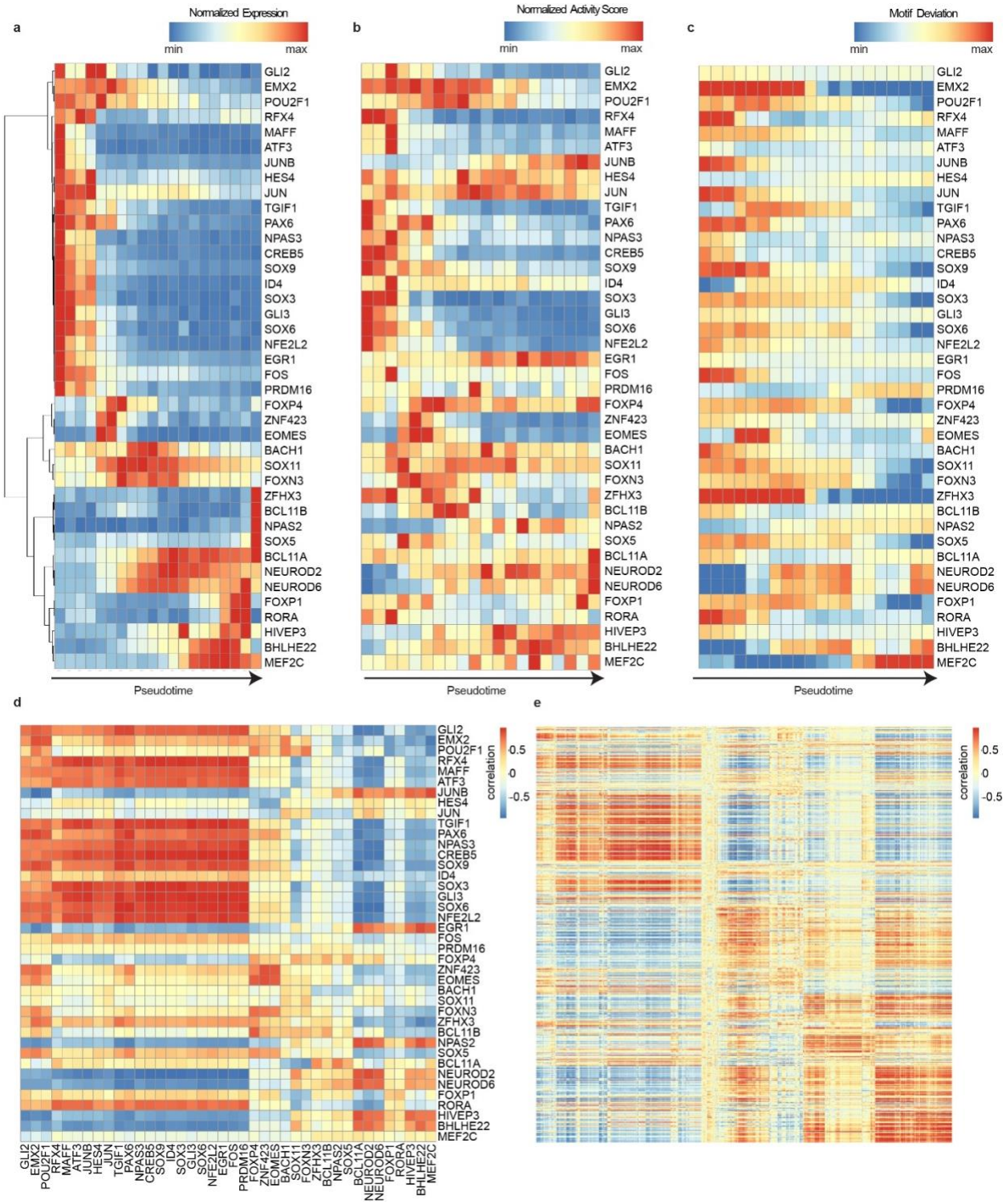
a
Normalized Expression
min
max
b
Normalized Activity Score
min
max
c
Motif Deviation
min
max
GLI2
EMX2
POU2F1
RFX4
MAFF
ATF3
JUNB
HES4
JUN
TGIF1
PAX6
NPAS3
CREB5
SOX9
ID4
SOX3
GLI3
SOX6
NFE2L2
EGR1
FOS
PRDM16
FOXP4
ZNF423
EOMES
BACH1
SOX11
FOXN3
ZFHX3
BCL11B
NPAS2
SOX5
BCL11A
NEUROD2
NEUROD6
FOXP1
RORA
HIVEP3
BHLHE22
MEF2C
Pseudotime
d
correlation
0.5
0
-0.5
e
correlation
0.5
0
-0.5

Figure S2.7: Dynamic patterns of gene expression, chromatin accessibility, and transcription factor motif enrichment across pseudotime. a) Heatmap of normalized expression of 40 key lineage transcription factors across pseudotime. Pseudotime is binned into 20 equally sized bins and accessibility is averaged across all cells in each bin. **b)** Heatmap of normalized gene activity scores for 40 key lineage transcription factors across pseudotime. Pseudotime is binned into 20 equally sized bins and activity scores are averaged across all cells in each bin. **c)** Heatmap of deviation scores of motif enrichments for 40 key lineage transcription factors across pseudotime. Pseudotime is binned into 20 equally sized bins and deviations are averaged across all cells in each bin. Deviation scores determined using ChromVAR[26] (Methods). **d)** Heatmap of correlations between gene expression and gene activity scores for 40 key lineage transcription factors. **e)** Heatmap of correlations between gene expression and gene activity scores for 615 cell type marker genes.

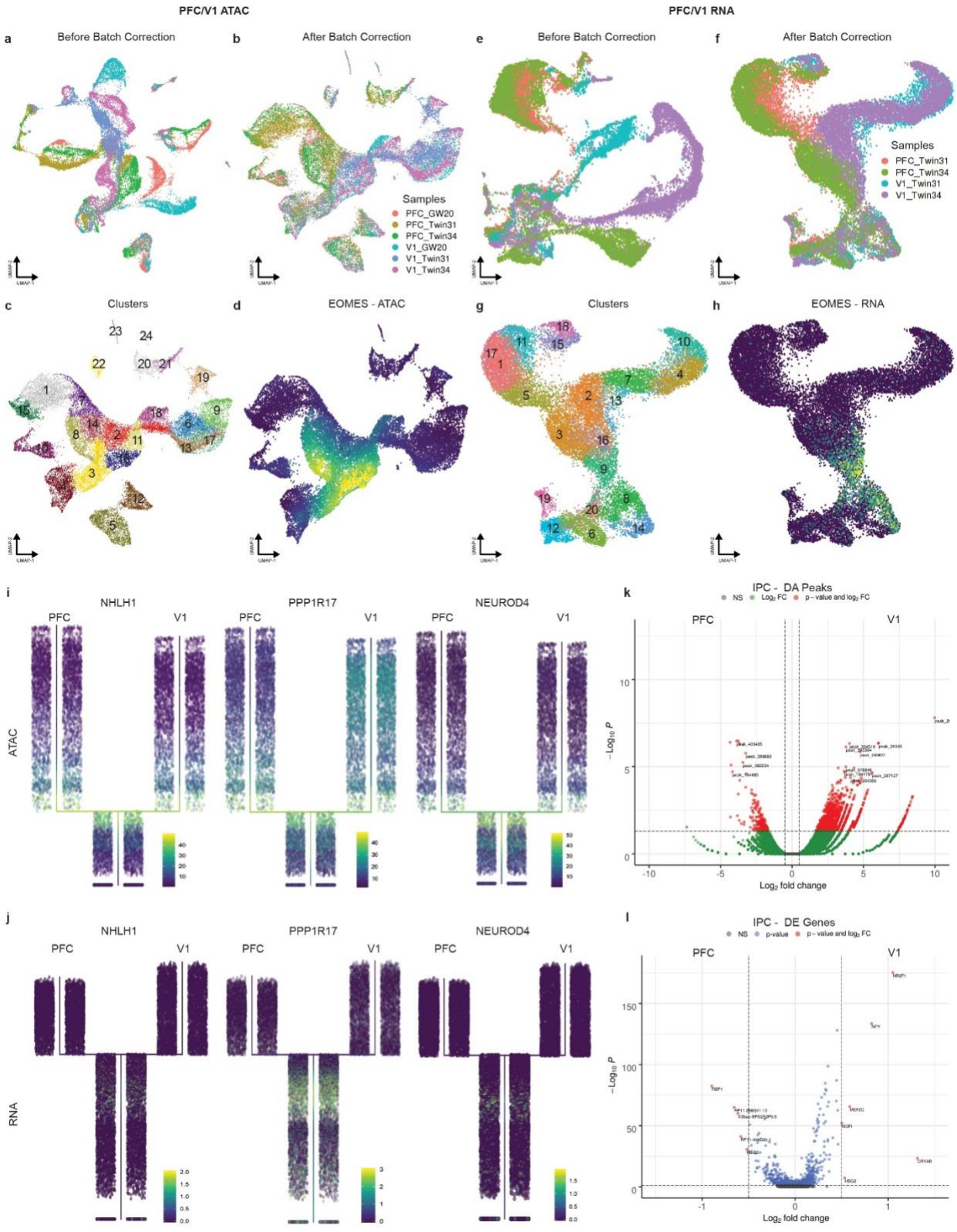
PFC/V1 ATAC
PFC/V1 RNA
a
Before Batch Correction
b
After Batch Correction
e
Before Batch Correction
f
After Batch Correction
Samples
PFC_GW20
PFC_Twin31
PFC_Twin34
V1_GW20
V1_Twin31
V1_Twin34
Samples
PFC_Twin31
PFC_Twin34
V1_Twin31
V1_Twin34
c
Clusters
d
EOMES - ATAC
g
Clusters
h
EOMES - RNA
i
NHLH1
PPP1R17
NEUROD4
PFC
V1
ATAC
k
IPC - DA Peaks
NS
Log2 FC
p - value and log2 FC
−Log10 P
Log2 fold change
j
NHLH1
PPP1R17
NEUROD4
PFC
V1
RNA
l
IPC - DE Genes
NS
p-value
p - value and log2 FC
−Log10 P
Log2 fold change

Figure S2.8: Chromatin state profiling reveals divergence of PFC and V1 excitatory lineages. a) UMAP projection of scATAC-seq cells from PFC and V1 samples before batch correction colored by sample. **b)** UMAP projection of scATAC-seq cells from PFC and V1 samples after batch correction colored by sample. **c)** UMAP projection of scATAC-seq cells from PFC and V1 samples colored by leiden cluster. **d)** UMAP projection of scATAC-seq cells from PFC and V1 samples colored by gene activity score for *EOMES*, a marker of IPCs. **e)** UMAP projection of scRNA-seq cells from PFC and V1 samples before batch correction colored by sample. **f)** UMAP projection of scRNA-seq cells from PFC and V1 samples after batch correction colored by sample. **g)** UMAP projection of scRNA-seq cells from PFC and V1 samples colored by leiden cluster. **h)** UMAP projection of scRNA-seq cells from PFC and V1 samples colored by expression of *EOMES*, a marker of IPCs. **i)** Projection of *NHLH1, PPP1R17*, and *NEUROD4* gene activity scores on PFC & V1 scATAC-seq cells ordered by pseudotime with PFC/V1 divergence branch point displayed. **j)** Projection of *NHLH1, PPP1R17*, and *NEUROD4* gene expression on PFC & V1 scRNA-seq cells ordered by pseudotime with PFC/V1 divergence branch point displayed. **k)** Volcano plot of peaks that are differentially accessible between PFC and V1 IPCs. Peaks highlighted in red have logFC > 0.5 and FDR < 0.05. (n = 1819). **l)** Volcano plot of differentially expressed genes between PFC and V1 IPCs. Genes highlighted in red have logFC > 0.5 and FDR < 0.05. (n = 11).

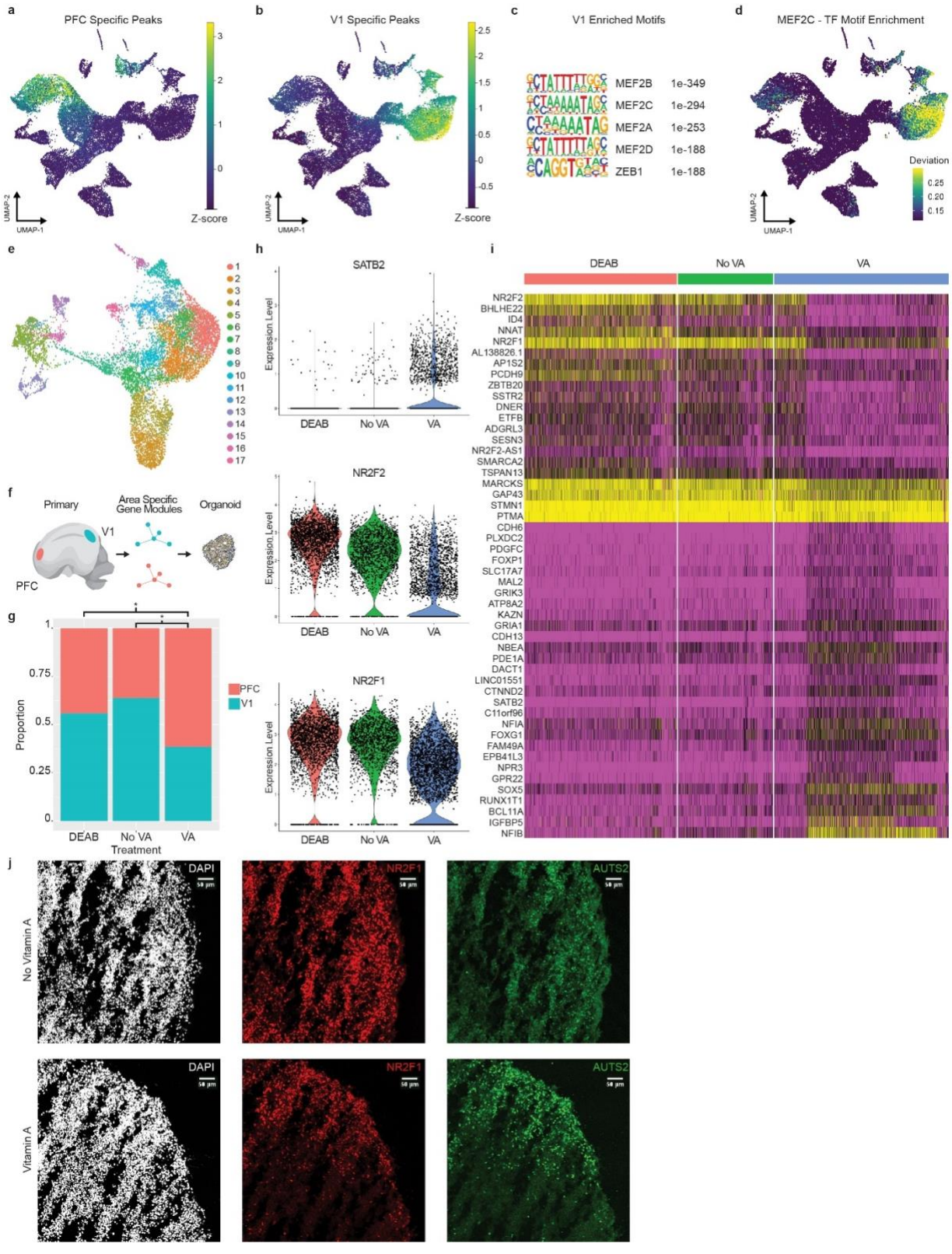
a
PFC Specific Peaks
Z-score
b
V1 Specific Peaks
Z-score
c
V1 Enriched Motifs
MEF2B 1e-349
MEF2C 1e-294
MEF2A 1e-253
MEF2D 1e-188
ZEB1 1e-188
d
MEF2C - TF Motif Enrichment
Deviation
UMAP-1
UMAP-2
e
f
Primary
Area Specific Gene Modules
Organoid
V1
PFC
g
Proportion
Treatment
DEAB
No VA
VA
PFC
V1
h
SATB2
NR2F2
NR2F1
Expression Level
i
DEAB
No VA
VA
NR2F2
BHLHE22
ID4
NNAT
NR2F1
AL138826.1
AP1S2
PCDH9
ZBTB20
SSTR2
DNER
ETFB
ADGRL3
SESN3
NR2F2-AS1
SMARCA2
TSPAN13
MARCKS
GAP43
STMN1
PTMA
CDH6
PLXDC2
PDGFC
FOXP1
SLC17A7
MAL2
GRIK3
ATP8A2
KAZN
GRIA1
CDH13
NBEA
PDE1A
DACT1
LINC01551
CTNND2
SATB2
C11orf96
NFIA
FOXG1
FAM49A
EPB41L3
NPR3
GPR22
SOX5
RUNX1T1
BCL11A
IGFBP5
NFIB
j
No Vitamin A
Vitamin A
DAPI
NR2F1
AUTS2
50 µm

Figure S2.9: Modeling the PFC/V1 split in the developing cortex. a) UMAP projection of Z-scores of enrichment of PFC specific peaks (n = 4,176) in all PFC & V1 scATAC-seq cells (Fisher's Exact, two-sided, FDR < 0.05). **b)** UMAP projection of Z-scores of enrichment of V1 specific peaks (n = 21,030) in all PFC & V1 scATAC-seq cells (Fisher's Exact, two-sided, FDR < 0.05). **c)** Top enriched transcription factor motifs in V1 specific peak set as determined by HOMER (Hypergeometric Test, one-sided). **d)** UMAP projection of ChromVAR deviation scores of motif enrichment of MEF2C in all PFC & V1 scATAC-seq cells. **e)** UMAP projection of scRNA-seq data from organoids (n=3) cultured in the presence of Vitamin A, without the presence of Vitamin A, and in the presence of DEAB. Cells colored by cluster. **f)** Schematic depiction of classifier method used to assign area identity to organoid cells based on defined area-specific gene modules. **g)** Barplot depicting proportion of excitatory neurons from each treatment group classified as more PFC-like or more V1-like based on calculation of module eigengene values for area-specific modules (Methods). Asterisks indicate significant differences in proportions (DEAB – PFC: 1160/2622, V1: 1462/2622; NoVA – PFC: 563/1556, V1: 993/1556; VA – PFC: 1831/2976, V1: 1145/2976) (Chi square test, one-sided, VA vs noVA: pval = 3.209e-59, VA vs DEAB: pval = 2.79e-38). **h)** Violin plots depicting expression levels of SATB2, NR2F1, and NR2F2 for excitatory neurons from each treatment group. **i)** Heatmap gene expression of differentially expressed genes between excitatory neurons cultured with and without Vitamin A. **j)** Images of organoids cultured with and without Vitamin A stained with DAPI and immunostained for NR2F1 and AUTS2. All images taken at 10x resolution. Representative images shown from n=3 lines.

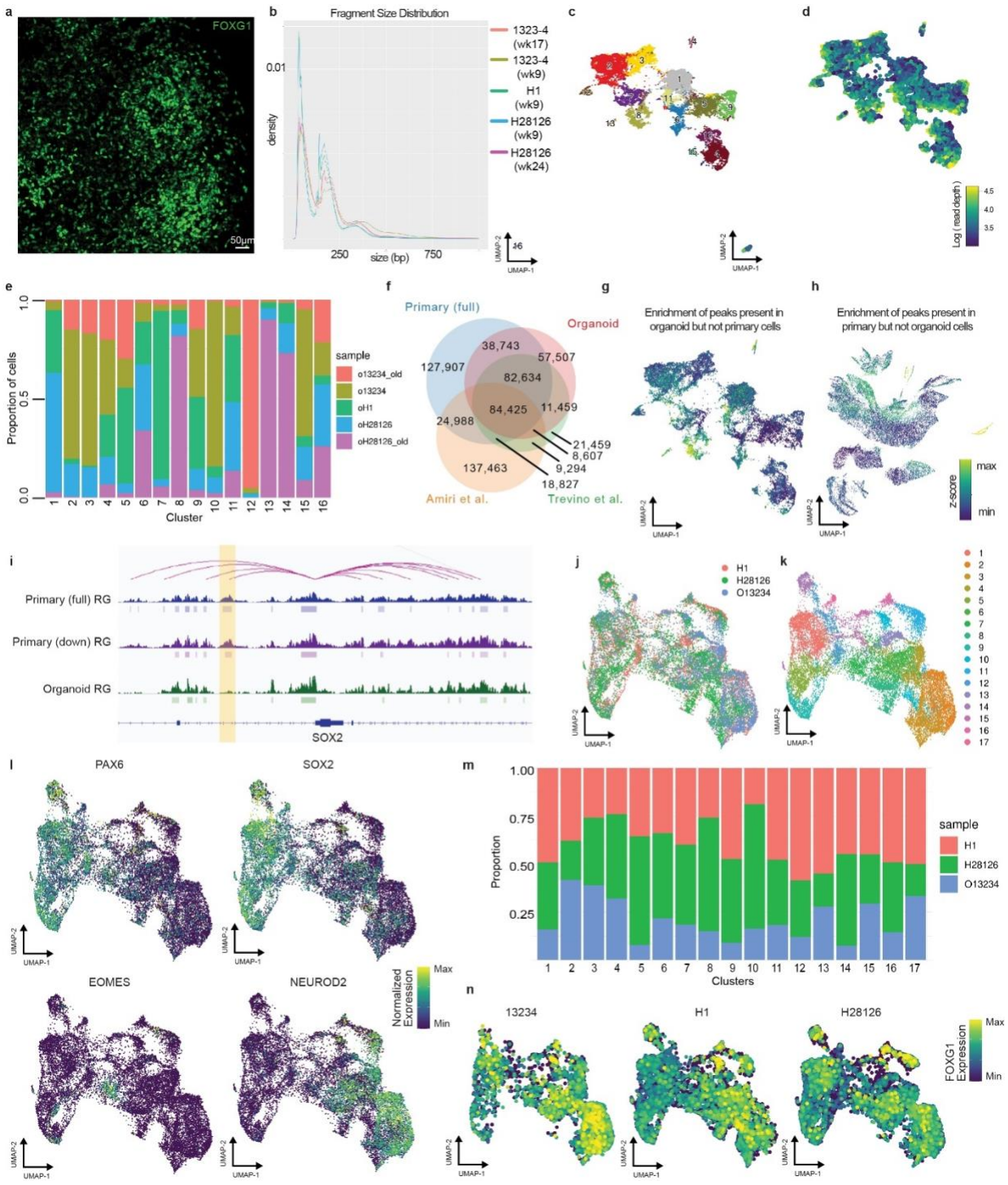
a
FOXG1
50μm
b
Fragment Size Distribution
density
size (bp)
1323-4 (wk17)
1323-4 (wk9)
H1 (wk9)
H28126 (wk9)
H28126 (wk24)
c
d
Log (read depth)
e
Proportion of cells
Cluster
sample
o13234_old
o13234
oH1
oH28126
oH28126_old
f
Primary (full)
Organoid
Amiri et al.
Trevino et al.
38,743
57,507
127,907
82,634
84,425
11,459
24,988
21,459
8,607
9,294
137,463
18,827
g
Enrichment of peaks present in organoid but not primary cells
h
Enrichment of peaks present in primary but not organoid cells
z-score
max
min
i
Primary (full) RG
Primary (down) RG
Organoid RG
SOX2
j
H1
H28126
O13234
k
l
PAX6
SOX2
EOMES
NEUROD2
Normalized Expression
Max
Min
m
Proportion
Clusters
sample
H1
H28126
O13234
n
13234
H1
H28126
FOXG1 Expression
Max
Min

Figure S2.10: Comparison of organoid and primary peaks reveal significant differences in the chromatin landscapes. a) Representative image of organoid (13234 line depicted) immunostained for FOXG1, a marker of cortical identity. Image taken at 10x resolution. Representative image shown from n=3 lines. **b)** Density curves of fragment size distributions for each organoid sample. Fragments are dedpulicated, uniquely mapped fragments that have passed quality filters (Methods). **c)** UMAP projection of all organoid scATAC-seq cells colored by cluster. **d)** UMAP projection of log(read depth). Fragments are dedpulicated, uniquely mapped fragments that have passed quality filters (Methods). **e)** Barplot depicting the proportions of cells in each cluster from each organoid sample. **f)** Venn diagram of overlaps of all primary scATAC-seq peaks, all organoid scATAC-seq peaks, all H3K27ac peaks from Amiri et al., 2018[11], and all ATAC-seq peaks from Trevino et al., 2020[49]. **g)** UMAP projection of Z-scores of enrichment of peaks that are present in the organoid scATAC-seq dataset but not in the primary scaTAC-seq dataset. **h)** UMAP projection of Z-scores of enrichment of peaks that are present in the primary scATAC-seq dataset but not in the organoid scaTAC-seq dataset. **i)** Genome browser tracks depicted a predicted enhancer of SOX2 that is present in both the full and down-sampled primary RGs but not detected in organoid RGs. **j)** UMAP projection of scRNA-seq data from the same three organoid lines used for scATAC-seq analysis colored by line (n = 19,509 cells). **k)** UMAP projection of scRNA-seq data colored by cluster. **l)** UMAP projection of normalized gene expression of *PAX6, SOX2, EOMES*, and *NEUROD2*. Maximum value was set at the 99th quantile. **m)** Barplot depicting the proportions of cells in each cluster from each organoid sample. **n)** UMAP projection of normalized gene expression of *FOXG1* split by organoid line. Maximum value was set at the 99th quantile.

REFERENCES

1. Nord AS, Blow MJ, Attanasio C, et al. Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell*. 2013;155(7):1521-1531. doi:10.1016/j.cell.2013.11.033
2. Nowakowski TJ, Bhaduri A, Pollen AA, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-)*. 2017;358(6368):1318-1323. doi:10.1126/science.aap8809
3. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*. 2020;180(3):568-584.e23. doi:10.1016/j.cell.2019.12.036
4. Hodge RD, Bakken TE, Miller JA, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*. 2019;573(7772):61-68. doi:10.1038/s41586-019-1506-7
5. Nowakowski TJ, Pollen AA, Sandoval-Espinosa C, Kriegstein AR. Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. *Neuron*. 2016;91(6):1219-1227. doi:10.1016/j.neuron.2016.09.005
6. Visel A, Taher L, Girgis H, et al. A high-resolution enhancer atlas of the developing telencephalon. *Cell*. 2013;152(4):895-908. doi:10.1016/j.cell.2012.12.041
7. Pattabiraman K, Golonzhka O, Lindtner S, et al. Article Transcriptional Regulation of Enhancers Active in Protodomains of the Developing Cerebral Cortex. *Neuron*. 2014;82(5):989-1003. doi:10.1016/j.neuron.2014.04.014

8. Mo A, Mukamel EA, Davis FP, et al. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain NeuroResource Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*. 2015;86(6):1369-1384. doi:10.1016/j.neuron.2015.05.018

9. de la Torre-Ubieta L, Stein JL, Won H, et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell*. 2018;172(1-2):289-295.e18. doi:10.1016/j.cell.2017.12.014

10. Luo C, Lancaster MA, Castanon R, Nery JR, Knoblich JA, Ecker JR. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep*. 2016;17(12):3369-3384. doi:10.1016/j.celrep.2016.12.001

11. Amiri A, Coppola G, Scuderi S, et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science (80-)*. 2018;362(6420):eaat6720. doi:10.1126/SCIENCE.AAT6720

12. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561):486-490. doi:10.1038/nature14590

13. Preissl S, Fang R, Huang H, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci*. 2018:1. doi:10.1038/s41593-018-0079-3

14. Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc B Biol Sci*. 2013;368(1632):20130025-20130025. doi:10.1098/rstb.2013.0025

15. Doan RN, Bae B Il, Cubelos B, et al. Mutations in Human Accelerated Regions

Disrupt Cognition and Social Behavior. *Cell*. 2016;167(2):341-354.e12. doi:10.1016/j.cell.2016.08.071

16. Johansen N, Quon G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol*. 2019;20(1):166. doi:10.1186/s13059-019-1766-4
17. Przytycki PF, Pollard KS. CellWalker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *bioRxiv*. January 2020:847657. doi:10.1101/847657
18. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330. doi:10.1038/nature14248
19. Won H, De La Torre-Ubieta L, Stein JL, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*. 2016;538(7626):523-527. doi:10.1038/nature19847
20. Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet*. 2019;51(12):1664-1669. doi:10.1038/s41588-019-0538-0
21. Song M, Pebworth M-P, Yang X, et al. Cell-type-specific 3D epigenomes in the developing human cortex. *Nature*. 2020. doi:10.1038/s41586-020-2825-4
22. Li M, Santpere G, Imamura Kawasawa Y, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science (80-)*. 2018;362(6420):eaat7615. doi:10.1126/science.aat7615
23. Markenscoff-Papadimitriou E, Whalen S, Przytycki P, et al. A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell*. 2020;0(0).

doi:10.1016/j.cell.2020.06.002

24. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the adult brain. *Science (80-)*. 2018;in press. doi:10.1126/science.aat8464

25. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res*. 2006;35(suppl_1):D88-D92. doi:10.1093/nar/gkl822

26. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14(10):975-978. doi:10.1038/nmeth.4401

27. Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421-427. doi:10.1038/nature13595

28. Turner TN, Coe BP, Dickel DE, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*. 2017;171(3):710-722.e12. doi:10.1016/j.cell.2017.08.047

29. Stessman HAF, Xiong B, Coe BP, et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat Genet*. 2017;49(4):515-526. doi:10.1038/ng.3792

30. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet*. 2014;46(10):1063-1071. doi:10.1038/ng.3092

31. Gandal MJ, Haney JR, Parikshak NN, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science (80-)*.

2018;359(6376):693-697. doi:10.1126/science.aad6469

32. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*. 2018;50(3):381-389. doi:10.1038/s41588-018-0059-2

33. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet*. 2019. doi:10.1038/s41588-019-0344-8

34. Howard DM, Adams MJ, Clarke T-K, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*. 2019;22(3):343-352. doi:10.1038/s41593-018-0326-7

35. Stahl EA, Breen G, Forstner AJ, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet*. 2019;51(5):793-803. doi:10.1038/s41588-019-0397-8

36. Zhu C, Yu M, Huang H, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol*. 2019;26(11):1063-1070. doi:10.1038/s41594-019-0323-x

37. Inoue F, Kreimer A, Ashuach T, Ahituv N, Yosef N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*. 2019;25(5):713-727.e10. doi:10.1016/j.stem.2019.09.010

38. Miller JA, Ding SL, Sunkin SM, et al. Transcriptional landscape of the prenatal human brain. *Nature*. 2014;508(7495):199-206. doi:10.1038/nature13185

39. Shibata M, Pattabiraman K, Lorente-Galdos B, et al. Regulation of Prefrontal Patterning, Connectivity and Synaptogenesis by Retinoic Acid. *bioRxiv*. January

2019:2019.12.31.891036. doi:10.1101/2019.12.31.891036

40. Molotkova N, Molotkov A, Duester G. Role of retinoic acid during forebrain development begins late when Raldh3 generates retinoic acid in the ventral subventricular zone. *Dev Biol*. 2007;303(2):601-610. doi:10.1016/j.ydbio.2006.11.035

41. Shabtai Y, Bendelac L, Jubran H, Hirschberg J, Fainsod A. Acetaldehyde inhibits retinoic acid biosynthesis to mediate alcohol teratogenicity. *Sci Rep*. 2018;8(1):347. doi:10.1038/s41598-017-18719-7

42. Bhaduri A, Andrews MG, Mancia Leon W, et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature*. 2020;578(7793):142-148. doi:10.1038/s41586-020-1962-0

43. Armentano M, Chou S-J, Srubek Tomassy G, Leingärtner A, O'Leary DDM, Studer M. COUP-TFI regulates the balance of cortical patterning between frontal/motor and sensory areas. *Nat Neurosci*. 2007;10(10):1277-1286. doi:10.1038/nn1958

44. Bedogni F, Hodge RD, Nelson BR, et al. Autism susceptibility candidate 2 (Auts2) encodes a nuclear protein expressed in developing brain regions implicated in autism neuropathology. *Gene Expr Patterns*. 2010;10(1):9-15. doi:10.1016/j.gep.2009.11.005

45. Cadwell CR, Bhaduri A, Mostajo-Radji MA, Keefe MG, Nowakowski TJ. Development and Arealization of the Cerebral Cortex. *Neuron*. 2019;103(6):980-1004. doi:10.1016/j.neuron.2019.07.009

46. Cederquist GY, Tchieu J, Callahan SJ, et al. A Multiplex Human Pluripotent Stem

Cell Platform Defines Molecular and Functional Subclasses of Autism-Related Genes. *Cell Stem Cell*. 2020;27(1):35-49.e6. doi:10.1016/j.stem.2020.06.004

47. Pollen AA, Bhaduri A, Andrews MG, et al. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell*. 2019;176(4):743-756.e17. doi:10.1016/j.cell.2019.01.017

48. Camp JG, Badsha F, Florio M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci*. 2015;112(51):15672 LP - 15677. doi:10.1073/pnas.1520760112

49. Trevino AE, Sinnott-Armstrong N, Andersen J, et al. Chromatin accessibility dynamics in a model of human forebrain development. *Science (80-)*. 2020;367(6476):eaay1645. doi:10.1126/science.aay1645

50. Burrows CK, Banovich NE, Pavlovic BJ, et al. Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLOS Genet*. 2016;12(1):e1005793. https://doi.org/10.1371/journal.pgen.1005793.

51. Lake BB, Chen S, Sos BC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol*. 2018;36(1):70-80. doi:10.1038/nbt.4038

52. Ramaswami G, Won H, Gandal MJ, et al. Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. *Nat Commun*. 2020;11(1):4873. doi:10.1038/s41467-020-18526-1

53. Skene NG, Bryois J, Bakken TE, et al. Genetic identification of brain cell types

underlying schizophrenia. *Nat Genet*. 2018;50(6):825-833. doi:10.1038/s41588-018-0129-5

54. van Neerven S, Kampmann E, Mey J. RAR/RXR and PPAR/RXR signaling in neurological and psychiatric diseases. *Prog Neurobiol*. 2008;85(4):433-451. doi:10.1016/j.pneurobio.2008.04.006

55. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*. 2017;14(10):959-962. doi:10.1038/nmeth.4396

56. Kadoshima T, Sakaguchi H, Nakano T, et al. Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell–derived neocortex. *Proc Natl Acad Sci*. 2013;110(50):20284 LP - 20289. doi:10.1073/pnas.1315710110

57. McGinnis C, Patterson D, Winkler J, et al. MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *bioRxiv*. 2018.

58. Kaya-Okur HS, Wu SJ, Codomo CA, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*. 2019;10(1):1930. doi:10.1038/s41467-019-09982-5

59. Fang R, Preissl S, Hou X, et al. Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv*. January 2019:615179. doi:10.1101/615179

60. van Dijk D, Sharma R, Nainys J, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*. 2018;174(3):716-729.e27.

doi:10.1016/j.cell.2018.05.061

61. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44(W1):W160-W165. doi:10.1093/nar/gkw257

62. Song M, Pebworth M-P, Yang X, et al. 3D Epigenomic Characterization Reveals Insights Into Gene Regulation and Lineage Specification During Corticogenesis. *bioRxiv*. January 2020:2020.02.24.963652. doi:10.1101/2020.02.24.963652

63. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576-589. doi:10.1016/j.molcel.2010.05.004

64. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*. 2015;31(14):2382-2383. doi:10.1093/bioinformatics/btv145

65. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 2010;28(5):495-501. doi:10.1038/nbt.1630

66. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496-502. doi:10.1038/s41586-019-0969-x

67. Pliner HA, Packer JS, McFaline-Figueroa JL, et al. Cicero Predicts cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*. 2018:1-14. doi:10.1016/j.molcel.2018.06.044

68. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis.

Science (80-). 2018;360(6392):eaar3131. doi:10.1126/science.aar3131

69. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst*. 2019;8(4):281-291.e9. doi:10.1016/j.cels.2018.11.005

70. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;20(1):296. doi:10.1186/s13059-019-1874-1

71. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420. doi:10.1038/nbt.4096

72. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559

73. Kaplanis J, Samocha KE, Wiel L, et al. Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. *bioRxiv*. January 2020:797787. doi:10.1101/797787

74. Finucane HK, Reshef YA, Anttila V, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet*. 2018;50(4):621-629. doi:10.1038/s41588-018-0081-4

75. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228-1235. doi:10.1038/ng.3404

CHAPTER 3

Conclusion

With the work described in this dissertation, we were able to generate one of the first single cell resolution epigenomic datasets for human cortex during prenatal developmental stages and use these data to make novel biological findings related to the process of cortical arealization. Using our dataset, we nominated thousands of cell type specific enhancer candidates active during human cortical development, we showed dynamic changes occur in the chromatin landscape that correspond with the process of neurogenesis, we benchmarked cerebral organoids as an epigenomic model for human cortical development, and we identified a novel role for retinoic acid in specification of prefrontal cortical identity. Excitingly, this finding linking increased retinoic acid signaling to prefrontal identity in the cortex was also found independently in a separate study around the time the manuscript represented in chapter 2 was submitted for publication[1]. This work merely represents the tip of the iceberg for how this dataset can be used to further our understanding of human cortical development and neurodevelopmental disease. I am hopeful that the research community will continue to use the data we generated to uncover deeper mechanistic understanding of the critical role that the chromatin landscape plays in the processes of cell fate specification and arealization and build better models for cortical development when primary tissues are inaccessible. Additionally, I am hopeful that this dataset will continue to be useful for those studying the genetic underpinning of neurodevelopmental disease. While we were able to find significant enrichment of neurodevelopmental disease-associated variants in predicted enhancers of excitatory and inhibitory neurons, better-powered genetic studies will enable

the use of single cell epigenomic data for further fine mapping of disease-associated variants to pinpoint the causal mutations, as seen in a recent study of neurodegenerative diseases[2].

As I write this dissertation, there is ongoing work in the Ahituv and Nowakowski labs to follow up on some of the findings from the chapter 2 manuscript. We are currently conducting massively parallel reporter assays (MPRAs)[3,4] in primary developing human cortical cells to characterize the enhancer activity of the cell type specific enhancer candidates we identified. Defining a set of enhancers with activity in discrete cell types will aid in the development of molecular tools that can achieve targeted cell type specific expression, enabling the study of specific cell types in a way that was previously unattainable[5,6]. In addition, we are testing the enhancer activity of candidates that intersect neurodevelopmental disease-associated variants to better understand the mechanism by which genetic variants may mediate disease via disruption of gene regulation during development. Lastly, we are following up on the finding that increased retinoic acid signaling plays a role in prefrontalization of the cortex with further experiments to understand the mechanism by which this fate specification occurs. In summary, though I will be moving on to the next phase of my career, I am hopeful that the modest contribution I have been able to make during my doctoral work will be useful to other scientists in the field and enable them to continue to unlock the secrets of the extraordinary human brain.

REFERENCES

1. Shibata M, Pattabiraman K, Lorente-Galdos B, et al. Regulation of Prefrontal Patterning, Connectivity and Synaptogenesis by Retinoic Acid. *bioRxiv*. January 2019:2019.12.31.891036. doi:10.1101/2019.12.31.891036

2. Corces MR, Shcherbina A, Kundu S, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet*. 2020;52(11):1158-1168. doi:10.1038/s41588-020-00721-x

3. Gordon MG, Inoue F, Martin B, et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat Protoc*. 2020;15(8):2387-2412. doi:10.1038/s41596-020-0333-5

4. Inoue F, Kreimer A, Ashuach T, Ahituv N, Yosef N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*. 2019;25(5):713-727.e10. doi:10.1016/j.stem.2019.09.010

5. Vormstein-Schneider D, Lin JD, Pelkey KA, et al. Viral manipulation of functionally distinct interneurons in mice, non-human primates and humans. *Nat Neurosci*. 2020;23(12):1629-1636. doi:10.1038/s41593-020-0692-9

6. Mich JK, Graybuck LT, Hess EE, et al. Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. *Cell Rep*. 2021;34(13). doi:10.1016/j.celrep.2021.108754

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:
Ryan Ziffra
A210D8C3D6DA44B...
Author Signature

8/25/2021
Date