

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Three-dimensional epigenomic characterization reveals insights into gene regulation and disease in the human brain

**Permalink**

<https://escholarship.org/uc/item/8tg7f8f5>

**Author**

Song, Michael

**Publication Date**

2020

Peer reviewed|Thesis/dissertation

Three-dimensional epigenomic characterization reveals insights into gene regulation and disease in the human brain

by  
Michael Song

DISSERTATION

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

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

*Yin Shen*

Yin Shen

AD18806FAF2442B...

Chair

DocuSigned by:

*Nadav Ahituv*

Nadav Ahituv

DocuSigned by:

*Hao Li*

Hao Li

C2CA172BE8684DD...

Committee Members



## Acknowledgements

First and foremost, I would like to thank my advisor Yin Shen for all the encouragement, support, opportunities, and insights she has provided for me during my time at UCSF. It's been a huge honor to work in her lab as her first graduate student, and there's no other environment I would have been so uniquely positioned for success. I would also like to express my gratitude to each of the lab members I've worked with over the years: Simon, Lenka, Jonghoon, Xingjie, Xiaoyu, Ian, David, Kirsty, Xiekui, Seth, Jerry, Asami, Bingkun, Niko, Lazaros, and Yien. The mini-science fiction book club that materialized in our bay was a ton of fun! Furthermore, I am immensely grateful to all my collaborators over the years: Mark-Phillip, Ming, Ivan, Jesse, Jia, Yun, Changxu, Mayank, Ting, Chao, Alex, Arnold, Amy, Fumitaka, Jialiang, Jeff, Sean, and countless more. Finally, I am hugely appreciative of the PSPG program, Deanna, Nadav, and Rebecca, for their confidence and patience in me, my thesis committee members Hao and Nadav, and my classmates for their support and interactions over the years. Outside the realm of science, I will always be grateful to my family and friends—all the adventures we've had and will continue to have over the years, the wonderful UCSF squash community and our amazing, scrappy team which participated in numerous battles and struggles at rival clubs across the city—a highlight being that epic playoff run with Ian, Jake, and David in 2019, and all of my past mentors for their guidance up to this point. Lastly, I where would I be without Ya-Ting, who's always been at my side, through thick and thin—here's to the next phase of our lives!

## Contributions

The following previously published materials have been incorporated into this thesis:

**Chapter 1.** Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I. R., Wang, C., Jacob, F., Wu, K., Traglia, M., Tam, T. W., Jamieson, K., Lu, S., Ming, G., Li, Y., Yao, J., Weiss, L. A., Dixon, J. R., Judge, L. M., Conklin, B. R., Song, H., Gan, L. Shen, Y. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nature Genetics*, 51 (8), 1252-1262.

**Chapter 2.** Song, M., Pebworth, M., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J. D., Choudhary, M., Cui, X., Jones, I. R., Bergenholtz, S., Eze, U. C., Juric, I., Li, B., Maliskova, L., Lee, J., Liu, W., Pollen, A. A., Li, Y., Wang, T., Hu, M., Kriegstein, A. R., Shen, Y. (2020). Cell type-specific 3D epigenomes in the developing human cortex. *Nature* (accepted).

# Three-dimensional epigenomic characterization reveals insights into gene regulation and disease in the human brain

Michael Song

## Abstract

Mutations in gene regulatory elements have been associated with a wide range of complex neuropsychiatric disorders. However, due to their cell type-specificity and difficulties in characterizing their regulatory targets, our ability to identify causal genetic variants has remained limited. To address these constraints, we perform integrative analysis of chromatin interactions, open chromatin regions, and transcriptomes for four neural cell types: iPSC-induced excitatory neurons, iPSC-derived hippocampal dentate gyrus (DG)-like neurons, iPSC-induced lower motor neurons, and primary astrocytes. Furthermore, we analyze radial glia (RG), intermediate progenitor cells (IPCs), excitatory neurons (eNs), and interneurons (iNs) isolated from mid-gestational human cortex samples. We utilize the chromatin interactions to link the promoters for genes participating in key neurological processes to distal regulatory elements such as enhancers and silencers that may be dysregulated in disease. In addition, we identify a novel chromatin feature which underlies cell type-specific transcription, which we term super interactive promoters (SIPs). Finally, we validate distal regulatory elements in both RG and eNs using CRISPR techniques as well as CRISPRview, a novel technique we developed for validating cell type-specific cis-regulatory elements in heterogeneous populations of primary cells. This work presents the first characterization of cell type-specific 3D epigenomes in these neurologically relevant cell types, advancing our knowledge of gene regulation during human brain development and complex neuropsychiatric disorders.

## Table of Contents

<b>Introduction.....</b>	<b>1</b>
<b>Chapter 1. Cis-Regulatory Chromatin Contacts in Neural Cells Reveal Contributions of Genetic Variants to Complex Disorders.....</b>	<b>4</b>
1.1. Profiling the epigenomic landscape of chromatin interactions in human neural cells.....	5
1.2. PIRs contribute to cellular identity. ....	11
1.3. Identification of regulatory targets for in vivo-validated enhancer elements. ....	16
1.4. Validation of PIRs in human neural cells using CRISPR techniques.....	17
1.5. Cell type-specific enrichment and regulatory target identification for neuropsychiatric disorder risk variants at PIRs.....	20
1.6. Validation of PIRs containing neuropsychiatric disorder risk variants. ....	26
1.7. Genetic variants contribute to chromatin interaction bias and changes in transcription. ....	28
1.8. Discussion. ....	30
1.9. Methods. ....	31
<b>Chapter 2. Cell Type-Specific 3D Epigenomes in the Developing Human Cortex.....</b>	<b>41</b>
2.1. Sorting specific cell types from the developing human cortex. ....	42
2.2. Characterization cell type-specific 3D epigenomes. ....	44
2.3. Chromatin interactions influence cell type-specific transcription. ....	46
2.4. Super interactive promoters are enriched for lineage-specific genes.....	49
2.5. Transposable elements in SIP formation.....	50
2.6. Developmental trajectories from RG to eNs. ....	54
2.7. Human-specific aspects of cortical development.....	55
2.8. Partitioning SNP heritability for complex disorders and traits. ....	56

<b>2.9. Characterizing distal interacting regions in primary cells.</b>	60
<b>2.10. Discussion.</b>	62
<b>2.11. Methods.</b>	64
 <b>Conclusion.</b>	 77
 <b>References.</b>	 79

## List of Figures

<b>Figure 1.1.</b> Overview of Hi-C technology. ....	2
<b>Figure 1.2.</b> Profiling of chromatin interactions in functionally distinct neural cell types.....	4
<b>Figure 1.3.</b> Immunofluorescence staining of cell type-specific markers.....	6
<b>Figure 1.4.</b> Transcriptional profiles are consistent with cellular identity. ....	7
<b>Figure 1.5.</b> Chromatin interactions link promoters and PIRs. ....	8
<b>Figure 1.6.</b> Integrative analysis of chromatin interactions, epigenomic features, and gene expression. ....	9
<b>Figure 1.7.</b> Enrichment of H3K27ac and CTCF binding sites at PIRs. ....	10
<b>Figure 1.8.</b> Chromatin interactions are cell type-specific. ....	11
<b>Figure 1.9.</b> Cell type-specific PIRs and TF motif enrichment analysis.....	12
<b>Figure 1.10.</b> GO enrichment analysis of cell type-specific PIRs. ....	13
<b>Figure 1.11.</b> Cell type-specific chromatin interactions with the <i>CHAT</i> and <i>TLR4</i> promoters. ....	14
<b>Figure 1.12.</b> Enrichment of in vivo-validated enhancer elements at PIRs. ....	17
<b>Figure 1.13.</b> Cell type-specific PIRs at the <i>CDK5RAP3</i> locus. ....	18
<b>Figure 1.14.</b> Validation of cell type-specific PIRs in neural cell types. ....	19
<b>Figure 1.15.</b> Genetic analysis of chromatin interactions with complex neuropsychiatric disorder-associated variants. ....	21
<b>Figure 1.16.</b> GO enrichment analysis of PIRs with complex neuropsychiatric disorder-associated variants. ....	23
<b>Figure 1.17.</b> SCZ-associated variants at the <i>FOXP1</i> locus. ....	24
<b>Figure 1.18.</b> Cell type-specific PIRs with putative regulatory SNPs.....	25
<b>Figure 1.19.</b> Validation of PIRs with complex neuropsychiatric disorder-associated variants. ..	27
<b>Figure 1.20.</b> Genetics variants contribute to chromatin interaction bias and alterations in gene expression. ....	29

<b>Figure 2.1.</b> Reproducibility between RNA-seq, ATAC-seq, and PLAC-seq replicates.....	43
<b>Figure 2.2.</b> Sorting specific cell types from the developing human cortex. ....	44
<b>Figure 2.3.</b> Features of 3D epigenomes during human corticogenesis. ....	45
<b>Figure 2.4.</b> GO enrichment analysis of cell type-specific chromatin interactions.....	46
<b>Figure 2.5.</b> H3K4me3-mediated chromatin interactions influence cell type-specific transcription. .....	47
<b>Figure 2.6.</b> SIPs are enriched for lineage-specific genes. ....	48
<b>Figure 2.7.</b> Features of SIPs. ....	50
<b>Figure 2.8.</b> Transposable elements in SIP formation.....	52
<b>Figure 2.9.</b> Validating ERVL-MaLR-localized ZNF143 motifs. ....	53
<b>Figure 2.10.</b> Developmental trajectories from RG to eNs. ....	54
<b>Figure 2.11.</b> Human-specific aspects of cortical development.....	56
<b>Figure 2.12.</b> Partitioning SNP heritability for complex neuropsychiatric disorders and traits.....	58
<b>Figure 2.13.</b> GO enrichment analysis of genes interacting with noncoding variants. ....	59
<b>Figure 2.14.</b> Overview of CRISPRview. ....	60
<b>Figure 2.15.</b> Validation of cell type-specific distal regulatory elements using CRISPRview. ....	61
<b>Figure 2.16.</b> Validation of additional RG-specific loci using CRISPRview. ....	62
<b>Figure 2.17.</b> Identifying significant H3K4me3-mediated chromatin interactions. ....	67

## List of Abbreviations

3C: chromosome conformation capture

AAVS1: adeno-associated virus integration site 1

AD: Alzheimer's disease

ADHD: attention deficit hyperactivity disorder

ASD: autism spectrum disorder

ALS: amyotrophic lateral sclerosis

APC: astrocyte progenitor cells

ATAC-seq: assay for transposable-accessible chromatin using sequencing

BH: Benjamini-Hochberg

BP: bipolar disorder

bp: base pairs

cDNA: complementary DNA

ChIP-seq: chromatin immunoprecipitation sequencing

CNS: central nervous system

CP: cortical plate

CRISPR: clustered regularly interspaced short palindromic repeats

CRISPRi: CRISPR interference

DG: dentate gyrus

eN: excitatory neuron

ENCODE: Encyclopedia of DNA Elements

EP: epilepsy

eQTL: expression quantitative trait loci

FACS: fluorescence-activated cell sorting

FDR: false discovery rate

FISH: fluorescent in situ hybridization

FPKM: fragments per kilobase million

FTD: frontotemporal dementia

GFAP: glial fibrillary acidic protein

GO: gene ontology

GW: gestational week

GWAS: genome wide association study

GZ: germinal zone

i<sup>3</sup>N iPSCs: WTC11 line containing an integrated, isogenic, and inducible *Ngn2* cassette

i<sup>3</sup>LMN iPSCs: WTC11 line containing an integrated, isogenic, and inducible *NGN2*, *ISL1*, and *LHX3* cassette

iN: interneuron

IPC: intermediate progenitor cell

iPSC: induced pluripotent stem cell

IQ: intelligence quotient

kb: kilobase

LD: linkage disequilibrium

Mb: megabase

MOI: multiplicity of infection

MP: mental process

Ngn2: neurogenin-2

pcHi-C: promoter capture Hi-C

PD: Parkinson's disease

PFA: paraformaldehyde

PIR: promoter-interacting regions

PLAC-seq: proximity ligation-assisted ChIP-seq

qPCR: quantitative polymerase chain reaction

RG: radial glia

RNA-seq: RNA sequencing

RPKM: reads per kilobase million

SCZ: schizophrenia

SEM: standard error of the mean

sgRNA: single guide RNA

SIP: super interactive promoter

SIPG: SIP group

SNP: single nucleotide polymorphism

TAD: topologically associated domains

TE: transposable elements

TF: transcription factor

TMM: trimmed mean of M-values

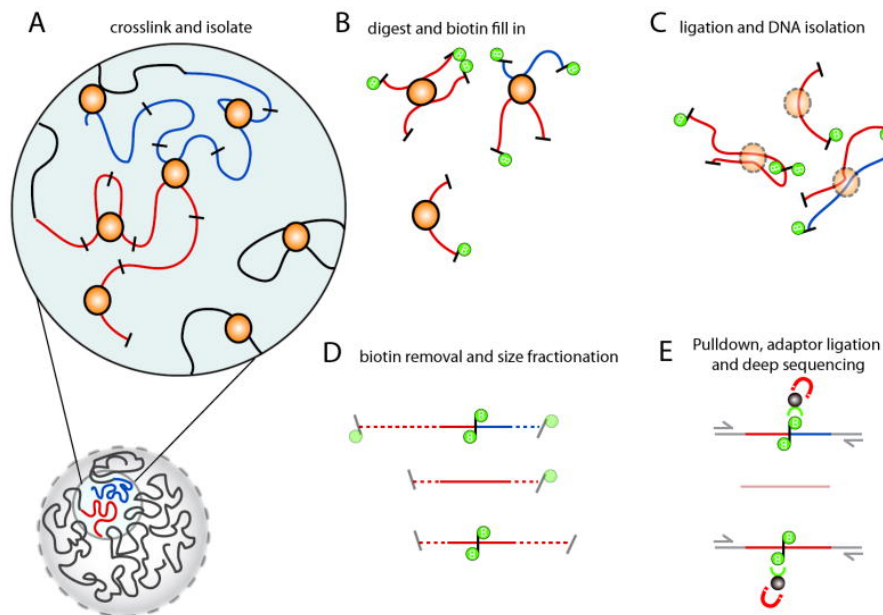
UD: unipolar depression

## Introduction

A large number of genetic variations and mutations which are associated with diverse human traits and diseases are located in putative regulatory regions. Genetic lesions in these regions can contribute to disease by modulating gene expression and disrupting finely tuned transcriptional networks in function and development. However, deciphering the roles of these regulatory variants in disease pathogenesis remains nontrivial due to their lack of annotation in the physiologically relevant cell types. In addition, regulatory elements often interact with their target genes over large genomic distances, precluding a straightforward mapping of regulatory element connectivity and limiting our interpretation of noncoding variants identified in genome wide association studies (GWAS). This is especially important for complex neuropsychiatric disorders such as Alzheimer's disease, autism spectrum disorder, and schizophrenia, each of which harbor highly heritable components which have eluded characterization to date (Chen et al., 2015; Rees et al., 2015). Typically, neighboring genes are assigned as risk loci for noncoding variants. However, this model is challenged by both experimental and computational evidence (Zhu et al., 2016; Mumbach et al., 2017). For instance, two independent obesity-associated SNPs in the *FTO* gene have been shown not to regulate *FTO*, but *IRX3* in the brain and both *IRX3* and *IRX5* in adipocytes, respectively (Smemo et al., 2014; Claussnitzer et al., 2015). The *FTO* locus in obesity illustrates the potentially intricate and cell type-specific manner in which noncoding variants contribute to disease. However, such well-annotated cases are rare, and we still lack systematic mappings of GWAS SNPs to their regulatory targets, especially in the context of complex neuropsychiatric disorders.

The recent emergence of high-throughput, genome-wide chromosome conformation capture (3C) technologies offers a promising approach for linking regulatory elements to their cognate genes. These technologies interrogate the complex three-dimensional structure of chromatin by fixing

protein-DNA complexes in place, digesting the DNA, then ligating the fragments such that interacting loci result in chimeric molecules that can be sequenced and analyzed (**Figure 1.1**). By interrogating the physical chromatin loops that bring regulatory elements such as enhancers and silencers in close spatial proximity with promoters to modulate transcription, detailed maps of regulatory contacts can be constructed for elucidating how transcriptional regulatory networks are established or dysregulated in the cell. Notably, these chromatin interactions are highly cell type-specific, with enhancer activity determined in part by transcription factor binding and remodeling of the local chromatin environment. For complex processes involving the highly orchestrated activities of multiple cell types as in development, the analysis of well-characterized, physiologically relevant cell lines or sorted cell populations offers significant advantages for identifying cell type-specific regulatory interactions which can then be validated in the appropriate experimental models.

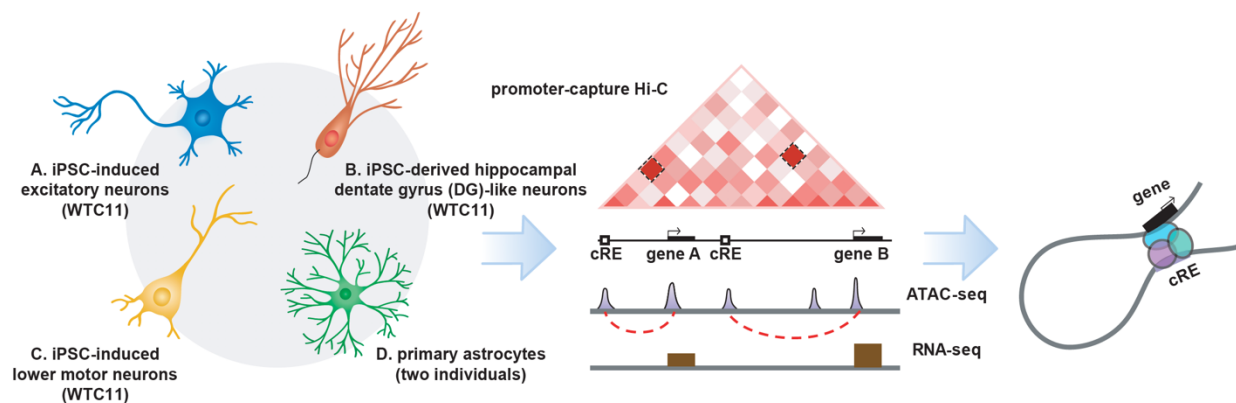


**Figure 1.1. Overview of Hi-C technology.** Hi-C detects chromatin interactions both within and between chromosomes by covalently crosslinking protein-DNA complexes using formaldehyde. Quantitation of chromatin interactions is achieved through high-throughput sequencing. (Adapted from Belton et al., 2012.)

In Chapter 1, we use wild type human iPSCs from the WTC11 line (Miyaoaka et al., 2014) to generate three functionally distinct neural cell types that are relevant to human brain development and disease: excitatory neurons (Wang et al., 2017), hippocampal dentate gyrus (DG)-like neurons (Mertens et al., 2015), and lower motor neurons (Fernandopulle et al., 2018). GFAP-positive astrocytes from the gastrulating brains of two individuals are also included in the study. By analyzing promoter-centric, long-range chromatin interactions using promoter capture Hi-C (pcHi-C), open chromatin regions using assay for transposase-accessible chromatin using sequencing (ATAC-seq), and transcriptomes using RNA sequencing (RNA-seq), we identify hundreds of thousands of cis-regulatory chromatin interactions between promoters and promoter-interacting regions (PIRs) in each cell type. We leverage these interactions to highlight putative gene targets for both in vivo-validated enhancer elements from the VISTA Enhancer Browser (Visel et al., 2007) and disease-associated variants (Buniello et al., 2019), enabling the elucidation and validation of PIRs driving diverse processes in cellular identity and disease. In Chapter 2, we adopt a similar approach, except we apply it to four major cell types in the developing human brain: radial glia (RG), intermediate progenitor cells (IPCs), excitatory neurons (eNs), and interneurons (iNs). We identify a novel chromatin feature which underlies cell type-specific transcription, which we term super interactive promoters (SIPs), and we explore the role of transposable elements (TEs) in their formation. Finally, we develop CRISPRview, a novel technique for validating cell type-specific cis-regulatory elements in heterogeneous populations of primary cells.

## Chapter 1. Cis-Regulatory Chromatin Contacts in Neural Cells Reveal Contributions of Genetic Variants to Complex Disorders

Previous epigenomic annotations of the germinal zone (GZ) and cortical plate (CP) in the developing human brain revealed the importance of chromatin structure in gene regulation and disease (Won et al., 2016; de la Torre-Ubieta et al., 2018). However, these studies used complex, heterogeneous tissues, limiting their abilities to interpret gene regulation in a cell type-specific manner. To address these constraints, we analyzed promoter-centric, long-range chromatin interactions using promoter capture Hi-C (pcHi-C), open chromatin regions using assay for transposase-accessible chromatin using sequencing (ATAC-seq), and transcriptomes using RNA sequencing (RNA-seq) for four neural cell types: iPSC-induced excitatory neurons, iPSC-derived hippocampal dentate gyrus (DG)-like neurons, iPSC-induced lower motor neurons, and primary astrocytes (**Figure 1.2**). We identify hundreds of thousands of cis-regulatory chromatin interactions between promoters and distal promoter-interacting regions (PIRs), enabling us to link regulatory elements to their cognate genes and identify processes that are putatively dysregulated in disease. Finally, we validate several PIRs using CRISPR techniques in human excitatory neurons, demonstrating that *CDK5RAP3*, *STRAP*, and *DRD2* are transcriptionally regulated by physically linked enhancers.

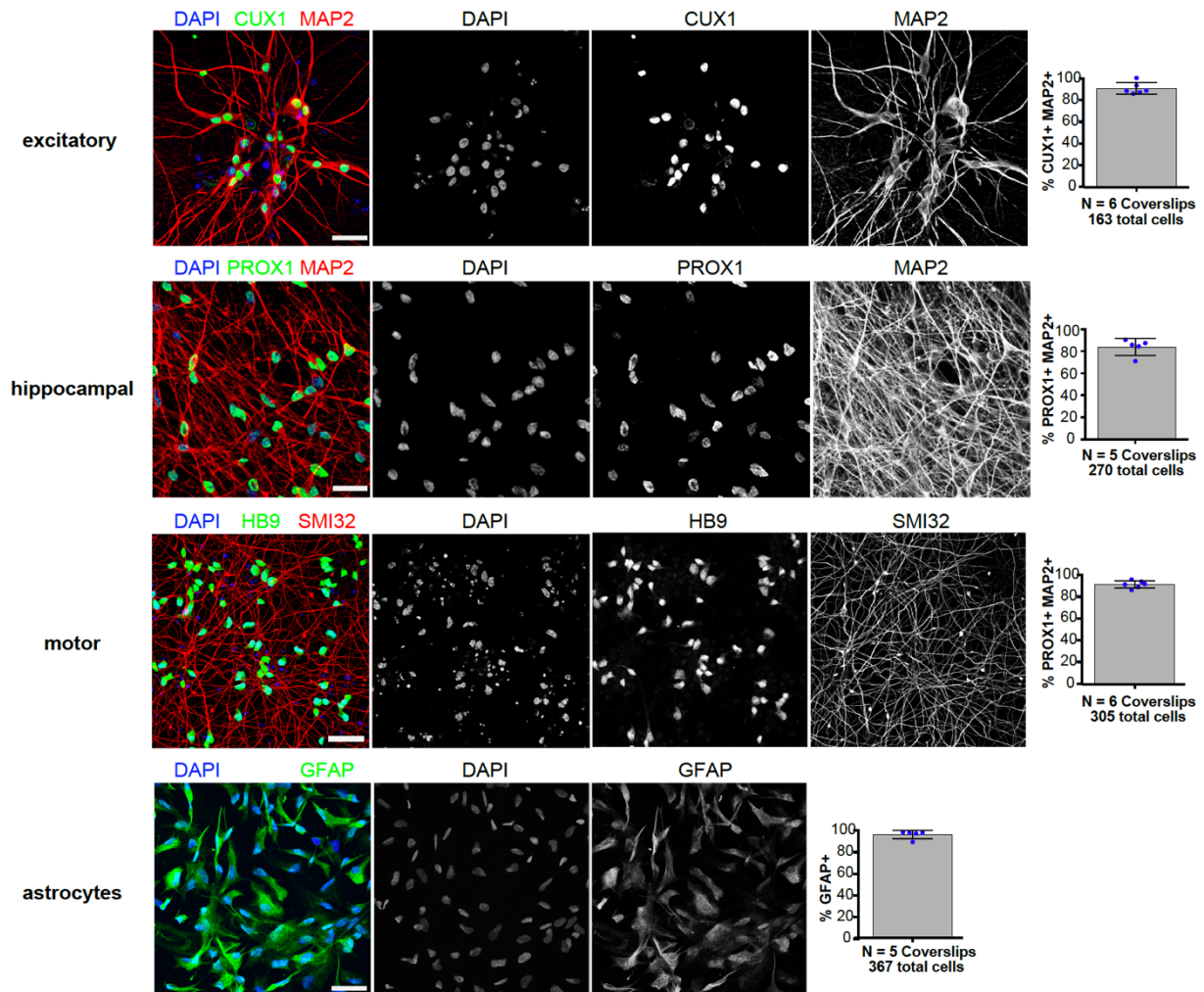


**Figure 1.2. Profiling of chromatin interactions in functionally distinct neural cell types.** Schematic of the study design for generating four functionally distinct neural cell types in the central nervous system (CNS) and performing integrative analysis of chromatin interactions using pcHi-C, open chromatin regions using ATAC-seq, and transcriptomes using RNA-seq.

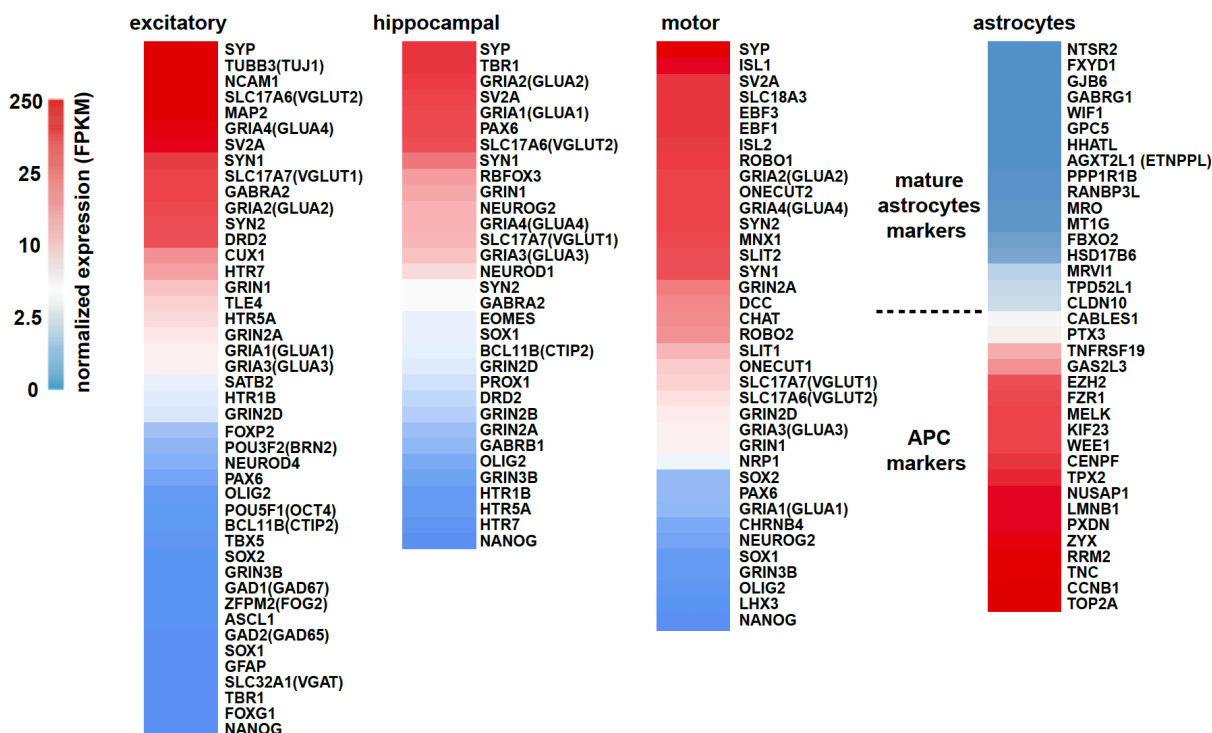
## 1.1 Profiling the epigenomic landscape of chromatin interactions in neural cells

To investigate general epigenomic features for cells in the human CNS, we focused on isogenic iPSC-induced excitatory neurons, iPSC-derived hippocampal dentate gyrus (DG)-like neurons, and iPSC-induced lower motor neurons, three neuronal subtypes which are currently impractical to isolate from primary tissue. Excitatory neurons were induced from a wild type male iPSC line (WTC11) containing an integrated, isogenic, and inducible neurogenin-2 (*Ngn2*) cassette (i<sup>3</sup>N iPSCs) with doxycycline-inducible *Ngn2* at the AAVS1 safe-harbor locus (Wang et al., 2017). The i<sup>3</sup>N iPSCs were used to prepare homogenous cultures of excitatory neurons expressing the glutamatergic neuron marker VGLUT1 and the cortical neuron marker CUX1 (Zhang et al., 2013; Wang et al., 2017), though *FOXG1* expression was not detected (**Figures 1.3 and 1.4**). Hippocampal DG-like neurons expressing the DG granule cell marker PROX1 were differentiated from a WTC11 line using factors as described previously (Yu et al., 2014; Mertens et al., 2015) (**Figures 1.3 and 1.4**). Finally, lower motor neurons were induced from a WTC11 line containing integrated, isogenic, and inducible *NGN2*, *ISL1*, and *LHX3* at the AAVS1 safe-harbor locus (i<sup>3</sup>LMN iPSCs) (Fernandopulle et al., 2018). The cells exhibited homogenous expression of the lower motor neuron markers HB9 and SMI32 (**Figures 1.3 and 1.4**). All three neuronal subtypes expressed the synaptic genes *SYN1* and *SYN2*, the NMDA receptor genes *GRIN1* and *GRIN2A*, and the AMPA receptor genes *GRIA1* and *GRIA2*, evidencing mature synaptic functions. We also included two batches of astrocytes isolated from 19 week gastrulating male fetal brain samples using GFAP as a selection marker. Astrocytes were cultured for two or fewer passages in vitro and confirmed for positive expression of GFAP prior to harvesting (**Figure 1.3**). Based on the age

of the donors and transcriptional signatures for dozens of marker genes distinguishing astrocyte progenitor cells (APCs) (e.g. *AGXT2L1* and *WIF1*) from mature astrocytes (e.g. *TOP2A* and *TNC*) (Zhang et al., 2016), the astrocytes were determined to be APCs (**Figure 1.4**).



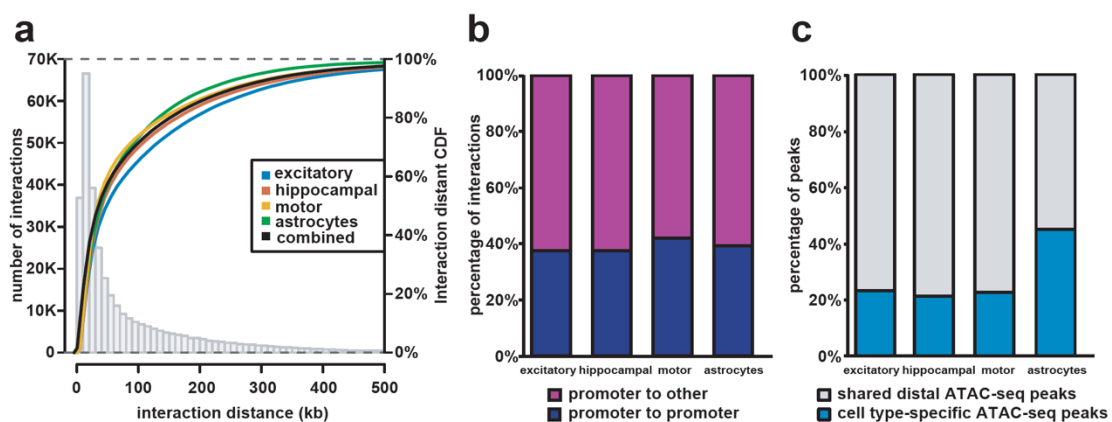
**Figure 1.3. Immunofluorescence staining of cell type-specific markers.** Immunofluorescence staining of key cell type-specific markers in excitatory neurons, hippocampal DG-like neurons, lower motor neurons, and astrocytes. Excitatory neurons were positively stained for CUX1, an upper cortical layer marker, and MAP2, a neuronal marker specifically expressed in dendrites. Hippocampal DG-like neurons were positively stained for PROX1, a transcription factor specifying granule cell identity in the DG. Lower motor neurons were positively stained for HB9, a motor neuron marker, and the pan-neuronal neurofilament marker SMI32. Astrocytes were positively stained for GFAP. The number of staining experiments and the total number of cells are indicated, and error bars represent the SEM. Scale bars indicate 25  $\mu$ m.



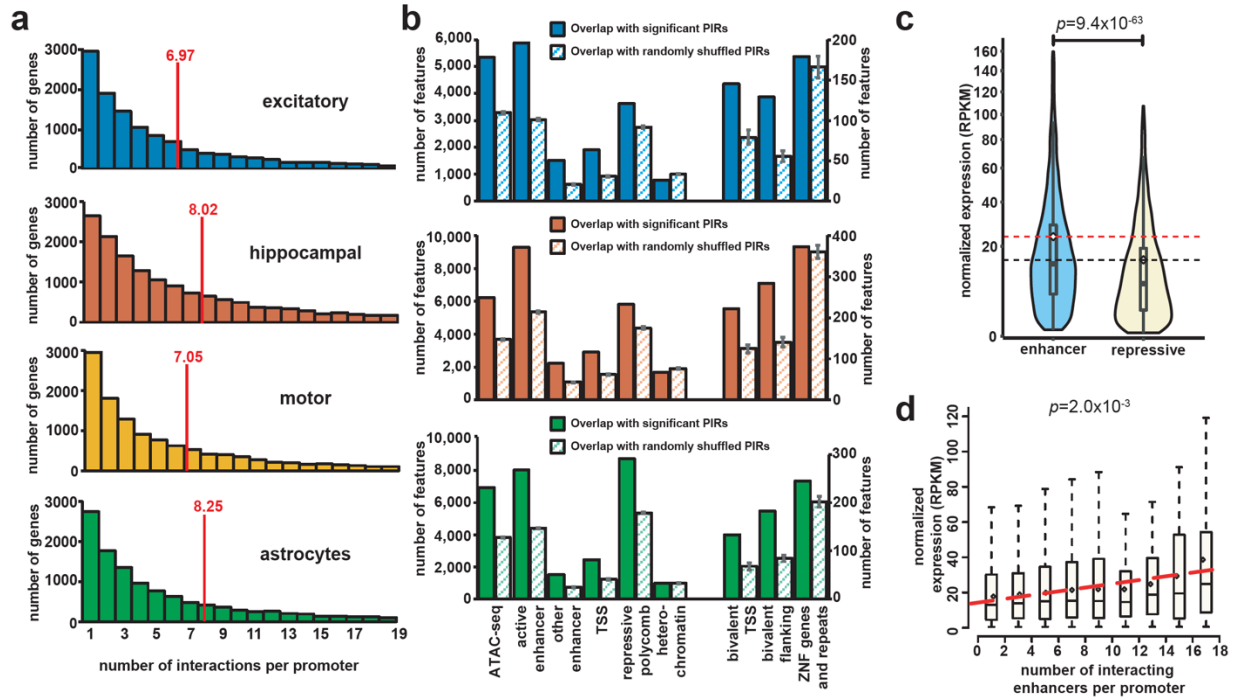
**Figure 1.4. Transcriptional profiles are consistent with cellular identity.** Heatmaps displaying gene expression results for key marker genes across the neural cell types. The astrocytes used in this study exhibit a gene expression profile consistent with APC identity.

We prepared pHi-C, ATAC-seq, and RNA-seq libraries using two to four biological replicates based on independent experiments for each cell type. Specifically, promoter-centric, long-range chromatin interactions were mapped using a set of 280,445 RNA probes targeting the promoters of 19,603 coding genes in GENCODE 19 (Jung et al., 2019). We first confirmed the reproducibility of contact frequency and saturation of inter-replicate correlation for our pHi-C libraries using HiCRep (Yang et al., 2017). Hierarchical clustering of ATAC-seq read density and gene expression similarly grouped the replicates by cell type, evidencing minimal variations during the cell derivation process. Using CHiCAGO (Cairns et al., 2016), we identified significant chromatin interactions with score  $\geq 5$  at 195,322 unique interacting loci across all four cell types, with 73,890, 108,156, 66,978, and 84,087 significant interactions being represented in the excitatory neurons,

hippocampal DG-like neurons, lower motor neurons, and astrocytes, respectively. Overall, 17,065 or 83.9% of coding gene promoters participate in interactions in at least one cell type, with 80% of PIRs interacting within a distance of 160 kb (**Figure 1.5a**). Over 97% of interactions occur within topologically associating domains (TADs) in human fetal brain tissues (Won et al., 2016). Furthermore, approximately 40% of interactions occur between promoter-containing bins, while 60% occur between promoter- and non-promoter-containing bins (**Figure 1.5b**). The observed counts of promoter-promoter interactions can potentially be attributed to transcriptional factories of coregulated genes, the widespread colocalization of promoters, and the capacity of many promoters to doubly function as enhancers (Jackson et al., 1993; Zhang et al., 2013; Engreitz et al., 2016; Diao et al., 2017). Finally, up to 40% of interacting distal open chromatin peaks are specific to each cell type (**Figure 1.5c**), suggesting that PIRs are capable of orchestrating cell type-specific gene regulation. Astrocytes in particular exhibit the highest proportion of cell type-specific open chromatin peaks, likely reflecting fundamental differences between the neuronal and glial lineages.



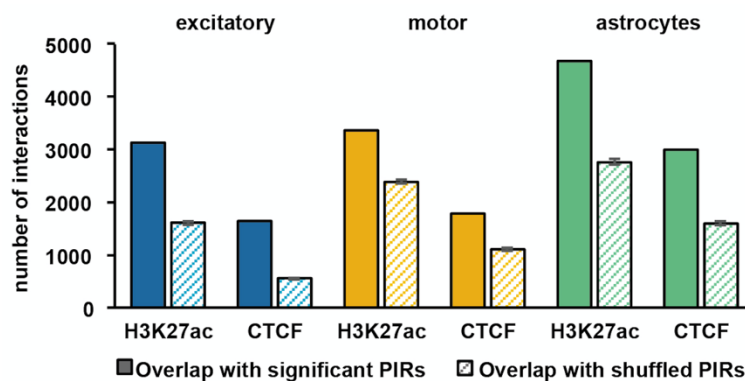
**Figure 1.5. Chromatin interactions link promoters and PIRs.** (a) Histogram and empirical CDF plots of interaction distances for each cell type. (b) Proportions of interactions between promoter-containing bins (blue) and promoter- and non-promoter-containing bins (purple) for each cell type. (c) Proportions of cell type-specific (blue) and shared (grey) distal open chromatin peaks at PIRs for each cell type.



**Figure 1.6. Integrative analysis of chromatin interactions, epigenomic features, and gene expression.** (a) Histograms of the number of PIRs interacting with each promoter for each cell type. Means are indicated. Only protein coding and noncoding RNA promoters interacting with at least one PIR are included. (b) Bar plots showing counts of epigenomic chromatin states inferred using ChromHMM in matched tissues overlapping significant (solid bars) versus randomly shuffled (striped bars) PIRs for each cell type. Means and the SEM for the number of overlaps across  $n=100$  sets of randomly shuffled PIRs are shown. (c) Comparative gene expression analysis across all cell types for expressed genes (normalized RPKM  $> 0.5$ ) whose promoters interact exclusively with either enhancer-PIRs ( $n=6,836$  genes) or repressive-PIRs ( $n=2,612$  genes) ( $P=9.4 \times 10^{-63}$ ,  $t=16.9$ ,  $df=6854.6$ , two-tailed two sample t-test). Boxplots indicate the median, IQR,  $Q1-1.5 \times IQR$ , and  $Q3+1.5 \times IQR$ . Means are indicated with dotted horizontal lines. (d) Distributions of gene expression values across all cell types for expressed genes (normalized RPKM  $> 0.5$ ) grouped according to the numbers of interactions their promoters form with enhancer-PIRs ( $P=2.1 \times 10^{-3}$ ,  $F_{1,7}=22.7$ , F-test for linear regression).

The majority of promoters interact with more than one PIR (Figure 1.6a), consistent with the large number of regulatory elements in the human genome and previous findings that promoters can be regulated by multiple enhancers (The ENCODE Project Consortium, 2012; Shen et al., 2012). To examine global chromatin signatures at PIRs, we leveraged chromatin states inferred by ChromHMM (Ernst et al., 2012) in matched human brain tissues from the Roadmap Epigenomics

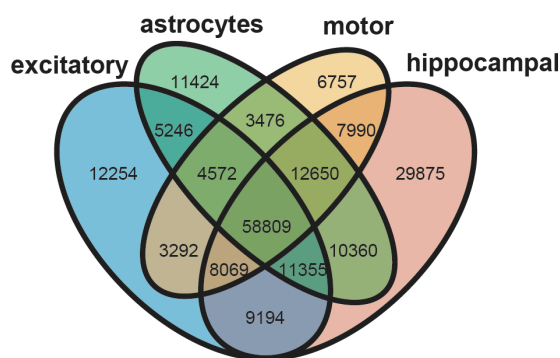
Project (dorsolateral prefrontal cortex for excitatory neurons, hippocampus middle for hippocampal DG-like neurons, and normal human astrocytes for astrocytes) (Roadmap Epigenomics Consortium, 2015). We show that PIRs are highly enriched for active chromatin features including open chromatin peaks, enhancers, and transcriptional start sites (TSSs) while simultaneously exhibiting depletion for repressive heterochromatin marks (**Figure 1.6b**). PIRs are also enriched for H3K27ac and CTCF binding sites mapped using CUT&RUN in excitatory and lower motor neurons and ChIP-seq in astrocytes from ENCODE (Skene et al., 2017; Davis et al., 2018) (**Figure 1.7**). Promoters interacting with enhancer-PIRs exhibit elevated levels of transcription compared to those interacting with repressive-PIRs ( $P=9.4 \times 10^{-63}$ , two-tailed two sample t-test) (**Figure 1.6c**). Multiple enhancer-PIRs also present evidence for additive effects on transcription. By grouping genes according to the number of interactions their promoters form with enhancer-PIRs, a modest correlation is observed between the number of interactions and the mean gene expression in each group ( $P=2.1 \times 10^{-3}$ , F-test for linear regression) (**Figure 1.6d**). Our results demonstrate that chromatin interactions identify PIRs which are not only enriched for regulatory features, but which are also capable of altering gene expression.



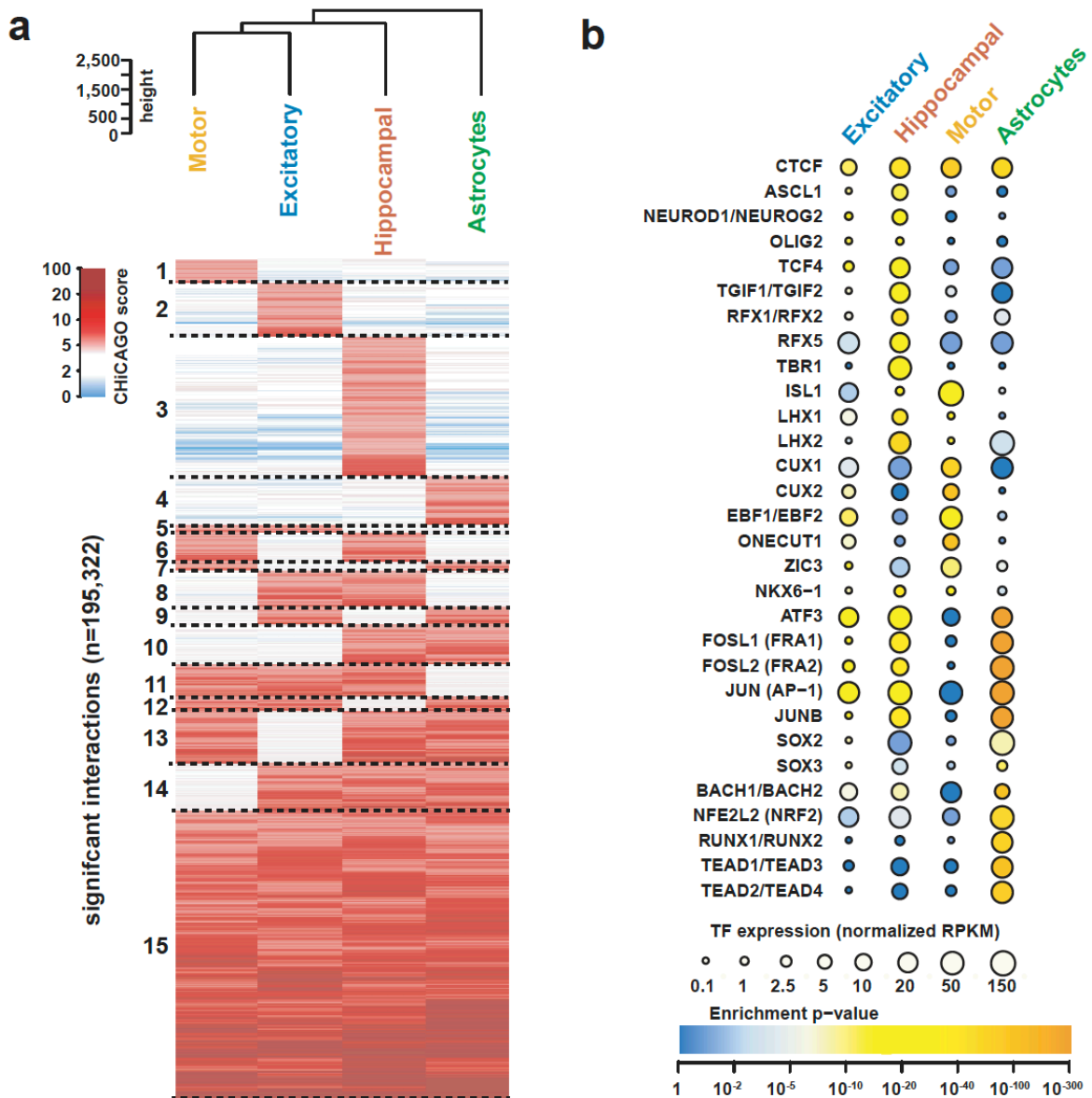
**Figure 1.7. Enrichment of H3K27ac and CTCF binding sites at PIRs.** Bar plots showing counts of H3K27ac and CTCF binding sites overlapping significant (solid bars) versus randomly shuffled (striped bars) PIRs for excitatory neurons, lower motor neurons, and astrocytes. Means and the SEM for the number of overlaps across  $n=100$  sets of randomly shuffled PIRs are shown.

## 1.2 PIRs contribute to cellular identity

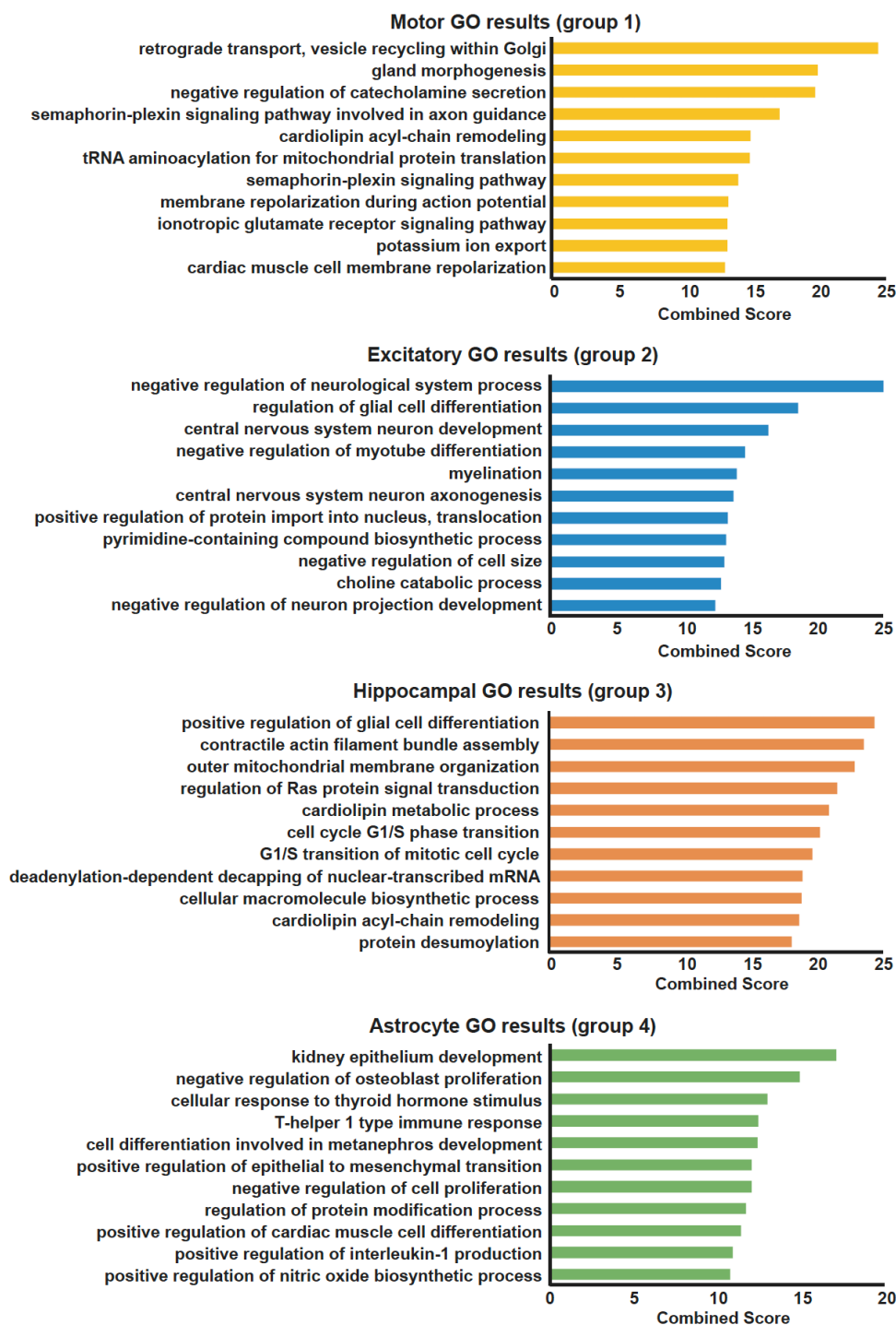
Chromatin interactions exhibit distinct patterns of cell type specificity, with tens of thousands of interactions observed to be specific for each cell type (**Figures 1.8** and **1.9a**). These interactions may underlie important functional differences between the cell types, with gene ontology (GO) enrichment analysis on genes interacting with cell type-specific PIRs yielding terms associated with neuronal function in the neuronal subtypes and immune function in the astrocytes (Kuleshov et al., 2016) (**Figure 1.10**). Meanwhile, 58,809 or 30.1% of unique interactions are shared across all four cell types, with neural precursor cell proliferation and neuroblast proliferation ranking among the top terms for genes participating in shared interactions. In conjunction with the enrichment of active chromatin signatures at PIRs, their association with cell type-specific processes suggests that PIRs harbor lineage-specific regulatory roles. Numerous promoters of differentially expressed genes including *OPHN1* in hippocampal DG-like neurons, *CHAT* in lower motor neurons, and *TLR4* in astrocytes form contacts with PIRs in their respective cell types (**Figure 1.11**). *OPHN1* stabilizes synaptic AMPA receptors and mediates long-term depression in the hippocampus, and its loss of function has been linked to mental retardation (Nadif Kasri et al., 2011). *CHAT* is a marker for lower motor neuron maturity and function, and *TLR4* is a key regulator of immune activation and synaptogenesis in astrocytes (Shen et al., 2016).



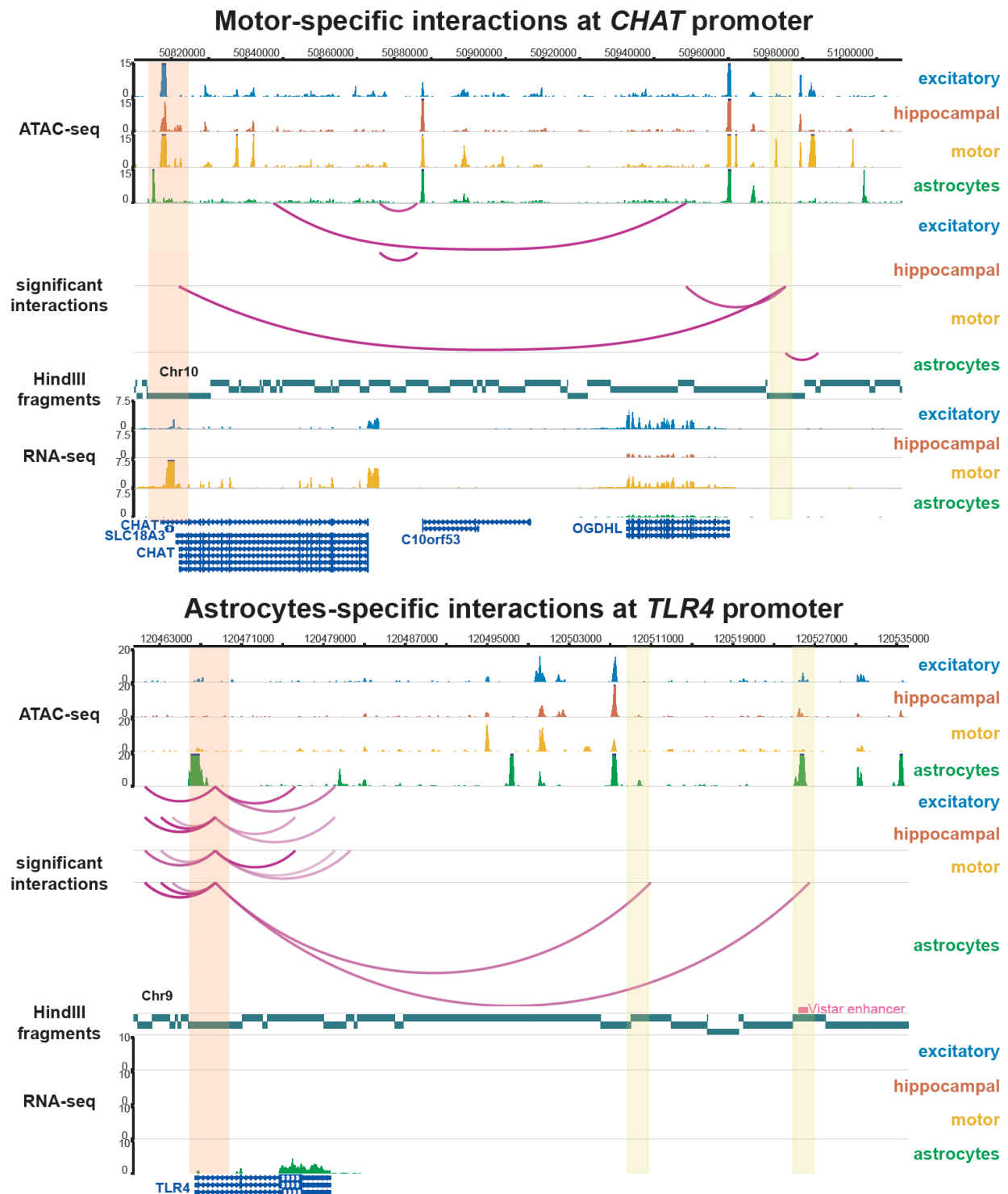
**Figure 1.8. Chromatin interactions are cell type-specific.** Venn diagram displaying counts of unique significant promoter-PIR interactions for excitatory neurons, hippocampal DG-like neurons, lower motor neurons, and astrocytes for each specificity pattern (groups 1-15 in **Figure 1.9a**).



**Figure 1.9. Cell type-specific PIRs and TF motif enrichment analysis.** (a) Classification of unique chromatin interactions with interaction score  $\geq 5$  in at least one cell type into specificity categories based on their scores in each cell type. Counts of significant promoter-PIR interactions in each specificity category are summarized in **Figure 1.8**. Cell types are hierarchically clustered based on their interaction scores over all interacting loci. (b) Enrichment of consensus TF motif sequences at open chromatin peaks in cell type-specific PIRs by motifs (rows) and cell types (columns). 1,145, 1,271, 843, and 2,566 peaks were used as input for the excitatory neurons, hippocampal DG-like neurons, lower motor neurons, and astrocytes, respectively. The color of each dot represents the degree of enrichment (calculated using the cumulative binomial distribution in HOMER) for each motif and cell type, and the size of each dot represents the gene expression (normalized RPKM) of the corresponding TFs for each motif. Entries with similar or identical consensus TF motif sequences are grouped for brevity.



**Figure 1.10. GO enrichment analysis of cell type-specific PIRs.** Top enriched GO terms from the “GO Biological Process 2018” ontology in Enrichr for genes participating in cell type-specific interactions with distal open chromatin peaks. Enriched GO terms are ranked by their combined scores (calculated by multiplying the log of the p-value based on Fischer’s exact test with the z-score of the deviation from the expected rank).

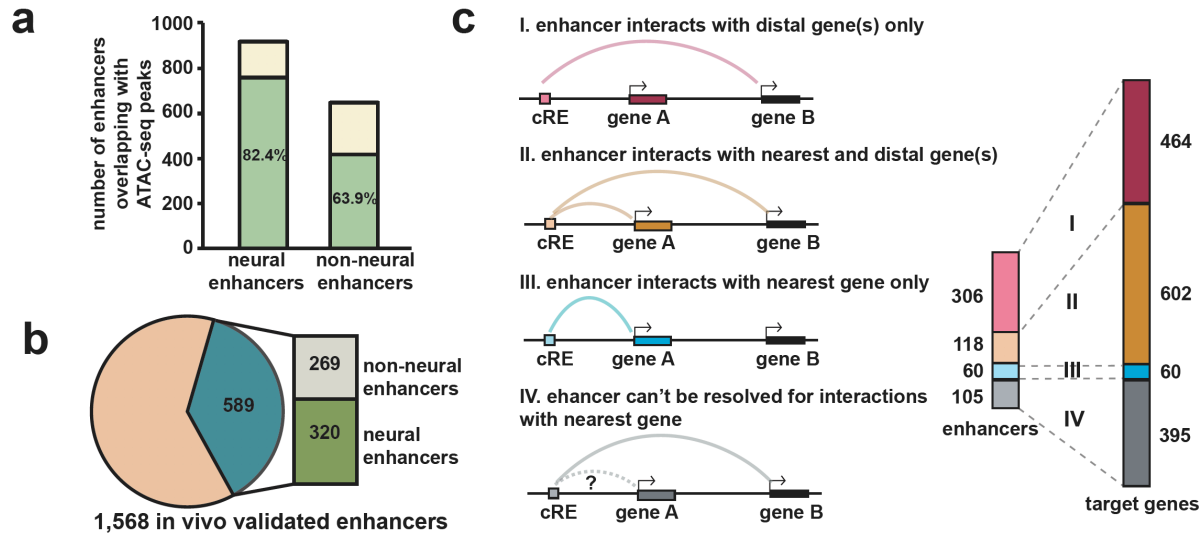


**Figure 1.11. Cell type-specific chromatin interactions with the *CHAT* and *TLR4* promoters.** Examples of cell type-specific interactions between PIRs (yellow) and the promoters of *CHAT* and *TLR4* (orange). ATAC-seq and RNA-seq signal are also displayed for each cell type.

Gene expression is coordinately controlled by transcription factors (TFs) and regulatory elements such as enhancers. Therefore, PIRs provide a novel perspective for investigating mechanisms underlying cell type-specific gene regulation. We used HOMER (Heinz et al., 2010) to evaluate TF motif enrichment at cell type-specific distal open chromatin peaks in PIRs for each cell type (**Figure 1.9b**). First, the CTCF motif is highly enriched across all cell types, consistent with its role in mediating looping within TADs (Guo et al., 2015; Handoko et al., 2011; Hou et al., 2010; Ren et al., 2017). Motifs for ASCL1, ISL1, NEUROG2, OLIG2, and ZIC3, TFs linked to neuronal fate commitment, are also broadly enriched across the neuronal subtypes. Additional TFs functioning in brain development include CUX1, CUX2, EBF1, EBF2, LHX1, LHX2, NKX6-1, ONECUT1, RFX1, RFX2, RFX5, TCF4, and TGIF2. The TBR1 motif is highly enriched in hippocampal DG-like neurons, consistent with its expression in the hippocampus (Chuang et al., 2014). Meanwhile, astrocytes are enriched for motifs in the Fos and Jun families, which contain key regulators for inflammatory and immune pathways. Also enriched in astrocytes are motifs for ATF3 and the RUNX and TEAD families, TFs with established roles in astrocyte differentiation, maturation, and proliferation. Motif enrichment is not always accompanied by expression of the corresponding TFs. This may reflect potential synergistic interactions between the cell types. For example, NRF2 is a key regulator of the oxidative stress response, and its activity has been shown to be repressed in neurons while also inducing a strong response in astrocytes (Liddell et al., 2017). Therefore, its shared expression may reflect the neuroprotective roles that astrocytes serve for other cell types. Alternatively, TFs do not have to be highly expressed to perform their cellular functions due to additional avenues for regulation at the post-transcriptional and post-translational levels. Our results demonstrate that PIRs contribute to cell fate commitment and are capable of identifying both known and novel regulators in a cell type-specific manner.

### 1.3 Identification of regulatory targets for in vivo-validated enhancer elements

Regulation of target genes by enhancers is thought to be mediated by physical chromatin looping. Congruent with this concept, chromatin interactions detected by pcHi-C can be used to link enhancers with their target genes. The VISTA Enhancer Browser is a database containing experimentally validated human and mouse noncoding sequences with enhancer activity (Visel et al., 2007). To date, it contains 2,956 tested sequences, 1,568 of which exhibit enhancer activity during embryonic development. However, the regulatory targets for these enhancer elements have remained largely uncharacterized. To address this knowledge gap, we provide cell type-specific annotations of putative target genes for each enhancer element using our chromatin interactions and open chromatin peaks. Across all cell types, our interactions recover 589 or 37.6% of positively tested enhancer elements with human sequences, 320 of which were further annotated as neural enhancers according to tissue-specific patterns of LacZ staining in mouse embryos (**Figures 1.12a-b**). Of the 589 interacting positive enhancer elements, 306 interact exclusively with 464 more distal genes (scenario I), 118 interact with both their nearest genes and 484 more distal genes (scenario II), and 60 interact exclusively with their nearest genes (scenario III) (**Figure 1.12c**). The remaining 105 elements could not be resolved for interactions with their nearest genes (scenario IV), though they interacted with 395 more distal genes. In total, our interactions identify 1,343 novel gene targets for positive enhancer elements in the VISTA Enhancer Browser, significantly expanding our knowledge of gene regulatory relationships at these loci.

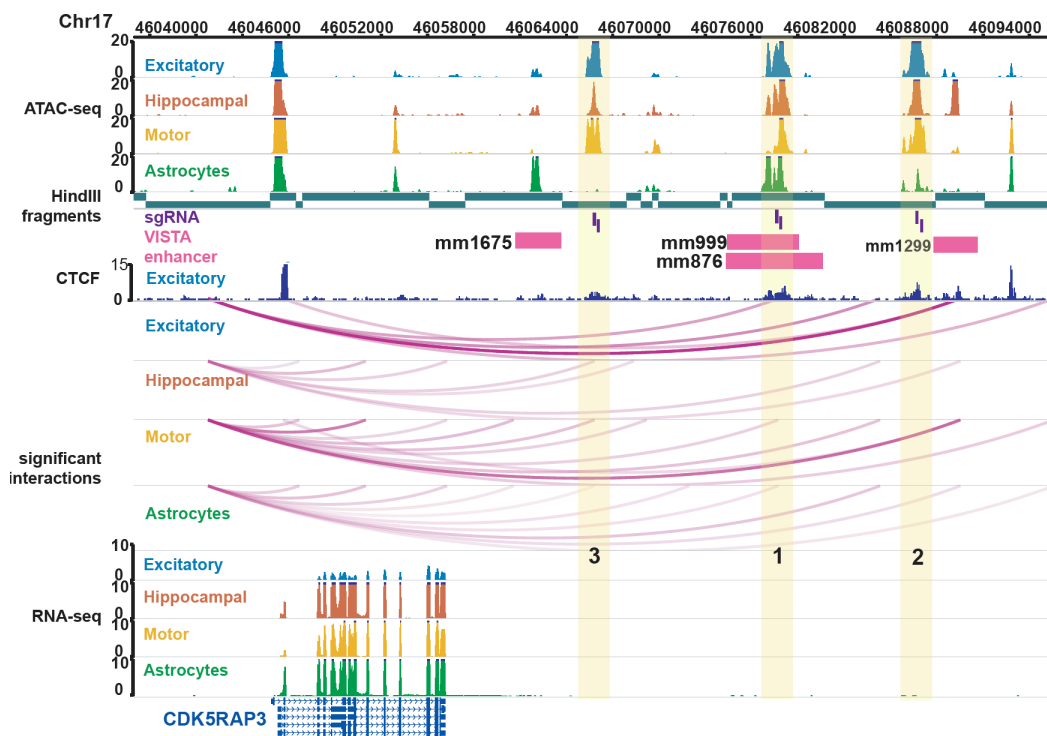


**Figure 1.12. Enrichment of in vivo-validated enhancer elements at PIRs.** (a) In vivo-validated enhancer elements with neural annotations overlap a significantly higher proportion of open chromatin peaks in the neural cells (757 of 919 elements) compared to enhancers with non-neural annotations (415 of 649 elements) ( $P < 2.2 \times 10^{-16}$ ,  $\chi^2=67.5$ ,  $df=1$ , Pearson's chi-squared test with Yates's correction). (b) Counts of in vivo-validated enhancer elements participating in chromatin interactions (589 of 1,568 elements) with neural and non-neural annotations. (c) Counts of interacting positive enhancer elements interacting exclusively with their nearest genes (blue), more distal genes (pink), or both (orange) and the number of target genes for each scenario (right).

#### 1.4 Validation of PIRs in human neural cells using CRISPR techniques

We validated two PIRs physically interacting up to 40 kb away with the *CDK5RAP3* promoter (regions 1 and 2) (Figure 1.13). *CDK5RAP3* regulates *CDK5*, which functions in neuronal development (Yin et al., 2005) and regulates proliferation in non-neuronal cells (Xie et al., 2003). Notably, both PIRs overlap open chromatin peaks as well as enhancers annotated with forebrain activity in the VISTA Enhancer Browser (mm876 and mm999 for region 1 and mm1299 for region 2) (Figure 1.14a). We targeted both regions for CRISPR deletion in the i<sup>3</sup>N iPSCs, followed by differentiation of the cells into excitatory neurons and quantification of any changes in gene expression by qPCR. Deleting the 2 kb open chromatin peak in region 1 led to significant

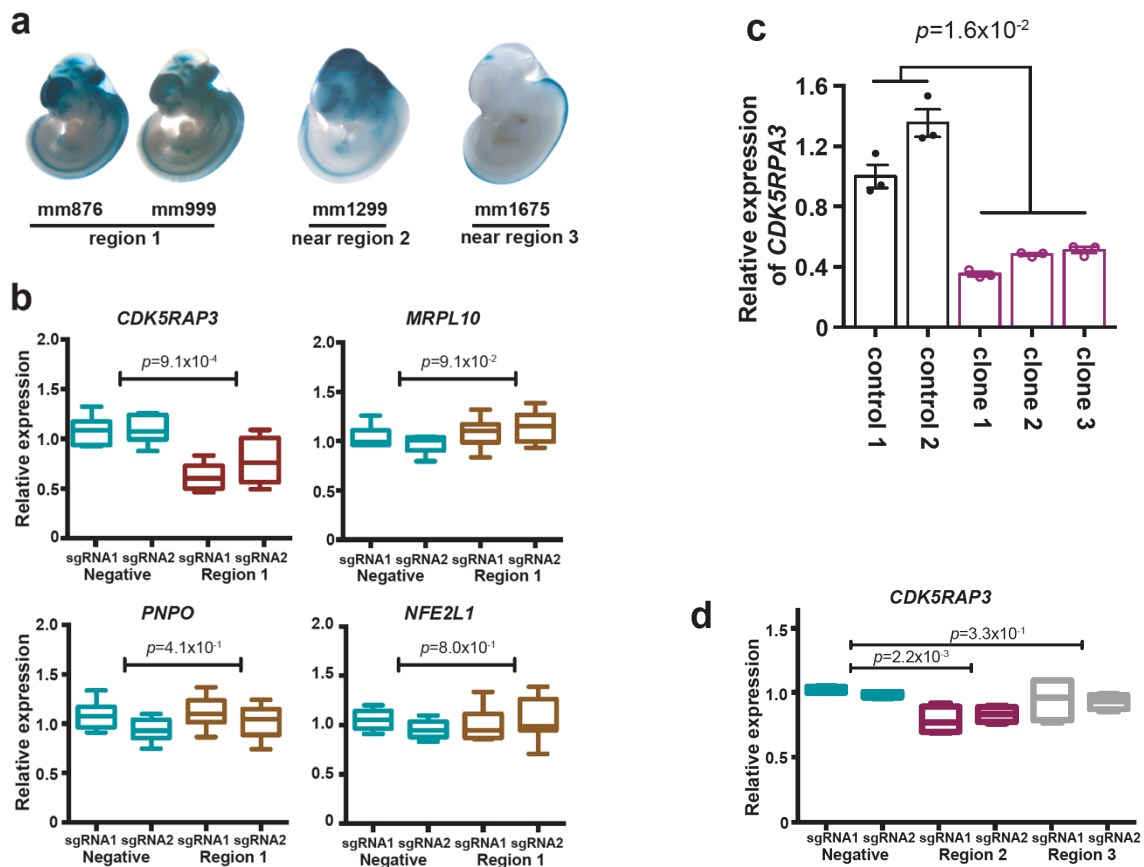
downregulation of *CDK5RAP3* expression across three independent clones (**Figure 1.14c**). However, upon trying to delete the open chromatin peak in region 2, we observed massive cell death of iPSCs immediately following introduction of the Cas9-sgRNA protein complex. We picked 48 individual clones from cells surviving the transfection, but failed to isolate any clones with deletions, suggesting that this locus is essential for maintaining *CDK5RAP3* expression and survival in iPSCs.



**Figure 1.13. Cell type-specific PIRs at the *CDK5RAP3* locus.** Open chromatin peaks in cell type-specific PIRs (regions 1, 2, and 3) interact with the *CDK5RAP3* promoter. Both in vivo-validated enhancer elements (pink) and CTCF binding sites in excitatory neurons (dark blue) are localized to all three regions.

To circumvent this lethal phenotype for iPSCs with region 2, we used CRISPR interference (CRISPRi) to silence both regions 1 and 2 in excitatory neurons. We also silenced a third region that interacts with the *CDKRAP3* promoter in the other cell types, but not in excitatory neurons

(region 3). We show that silencing of regions 1 and 2, but not region 3, leads to significant downregulation of *CDK5RAP3* expression without influencing the expression of nearby genes (Figures 1.14b and 1.14d). Interestingly, an enhancer with spinal cord activity proximal to region 3 (mm1675) interacts with the *CDK5RAP3* promoter in lower motor neurons and astrocytes, but not in the other cell types. Overall, these results demonstrate that chromatin interactions recapitulate cell type-specific patterns of enhancer activity, underscoring the importance of studying epigenomic regulation in the appropriate cell types.



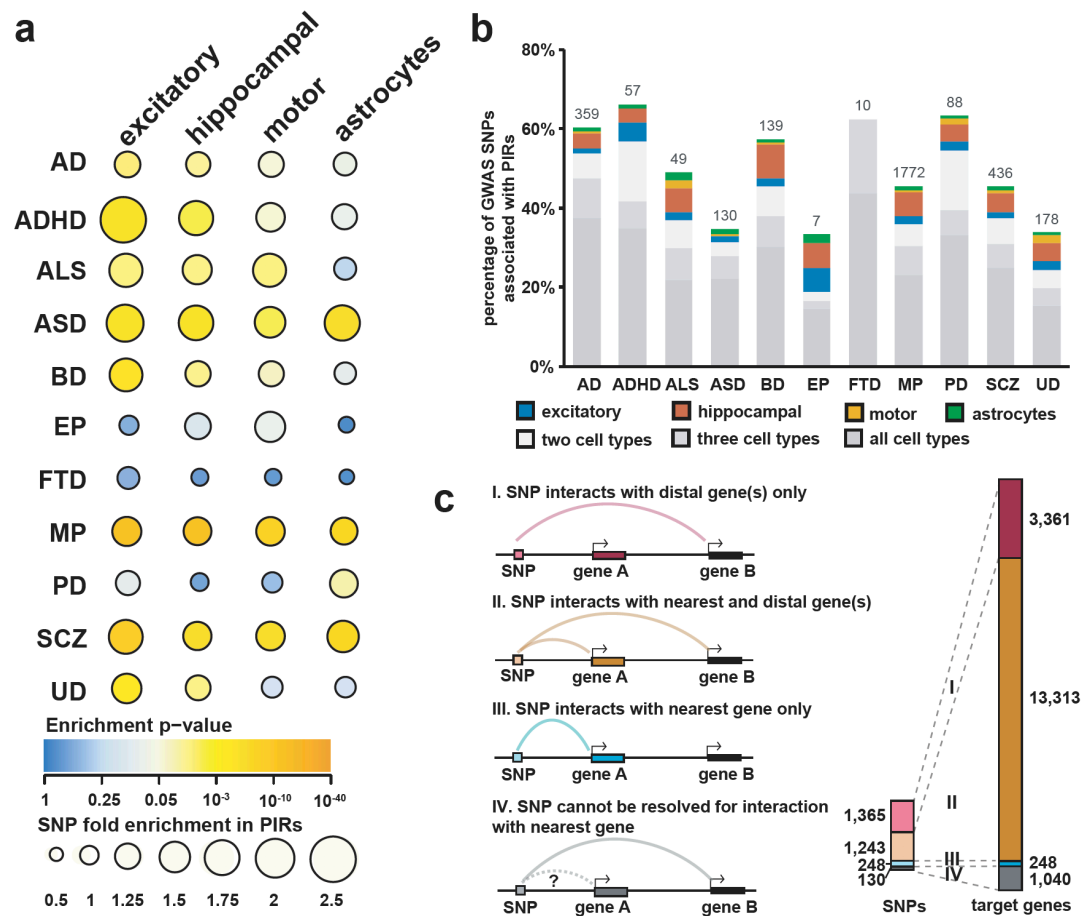
**Figure 1.14. Validation of cell type-specific PIRs in neural cell types.** (a) LacZ staining in mouse embryos reveals tissue-specific patterns of enhancer activity. (b) CRISPRi silencing of region 1 results in significant downregulation of *CDK5RAP3* expression in excitatory neurons ( $P=9.1 \times 10^{-4}$ ,  $t=4.65$ ,  $df=10$  two tailed two sample t-test). The neighboring genes *MRPL10*, *PNPO*, and *NFE2L1* were unaffected ( $P=9.1 \times 10^{-2}$ ,  $t=1.87$ ,  $df=10$ ,  $P=4.1 \times 10^{-1}$ ,  $t=0.853$ ,  $df=10$ , and  $P=8.0 \times 10^{-1}$ ,  $t=0.259$ ,  $df=10$  respectively, two-tailed two sample t-test). Three independent

replicates per condition and two sgRNAs per replicate were used for each experiment. Boxplots indicate the median, IQR, minimum, and maximum. **(c)** Significant downregulation of *CDK5RAP3* expression was observed across three independent clones containing biallelic deletions for the PIR in excitatory neurons ( $P=1.6 \times 10^{-2}$ ,  $t=4.98$ ,  $df=3$ , two-tailed two sample t-test). Error bars represent the SEM. **(d)** CRISPRi silencing of region 2, but not region 3, results in significant downregulation of *CDK5RAP3* expression in excitatory neurons ( $P=2.2 \times 10^{-3}$ ,  $t=5.11$ ,  $df=6$  and  $P=3.3 \times 10^{-1}$ ,  $t=1.05$ ,  $df=6$  respectively, two-tailed two sample t-test). Two independent replicates per condition and two sgRNAs per replicate were used for each experiment.

## **1.5 Cell type-specific enrichment and regulatory target identification for neuropsychiatric disorder risk variants at PIRs**

Previous large-scale epigenomic studies of human tissues and cell lines highlighted the importance of disease-associated variants at distal regulatory regions and the need for high-throughput approaches to prioritize variants for further validation (Roadmap Epigenomics Consortium, 2015). Therefore, we used our chromatin interactions to annotate complex neuropsychiatric disorder- and trait-associated variants from the GWAS Catalog (Buniello et al., 2019). We mined a total of 6,396 unique GWAS SNPs at a significance threshold of  $10^{-6}$  for eleven traits including Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), amyotrophic lateral sclerosis (ALS), bipolar disorder (BD), epilepsy (EP), frontotemporal dementia (FTD), mental process (MP), Parkinson's disease (PD), schizophrenia (SCZ), and unipolar depression (UD). We identified linked SNPs at a linkage disequilibrium (LD) threshold of 0.8 using HaploReg (Ward et al., 2012) for a total of 95,954 unique SNPs across all traits. We find that SNPs are enriched at PIRs in a disease- and cell type-specific manner (**Figure 1.15a**), with ASD, MP, and SCZ SNPs enriched at PIRs across all cell types. UD SNPs are enriched exclusively in excitatory and hippocampal DG-like neurons, whereas AD, ADHD, and BD SNPs also exhibit enrichment in lower motor neurons. ALS SNPs are enriched in the neuronal subtypes but not in astrocytes, consistent with the characterization of ALS as a motor neuron disease and reinforcing evidence for its role in hippocampal degeneration (Abdulla et al.,

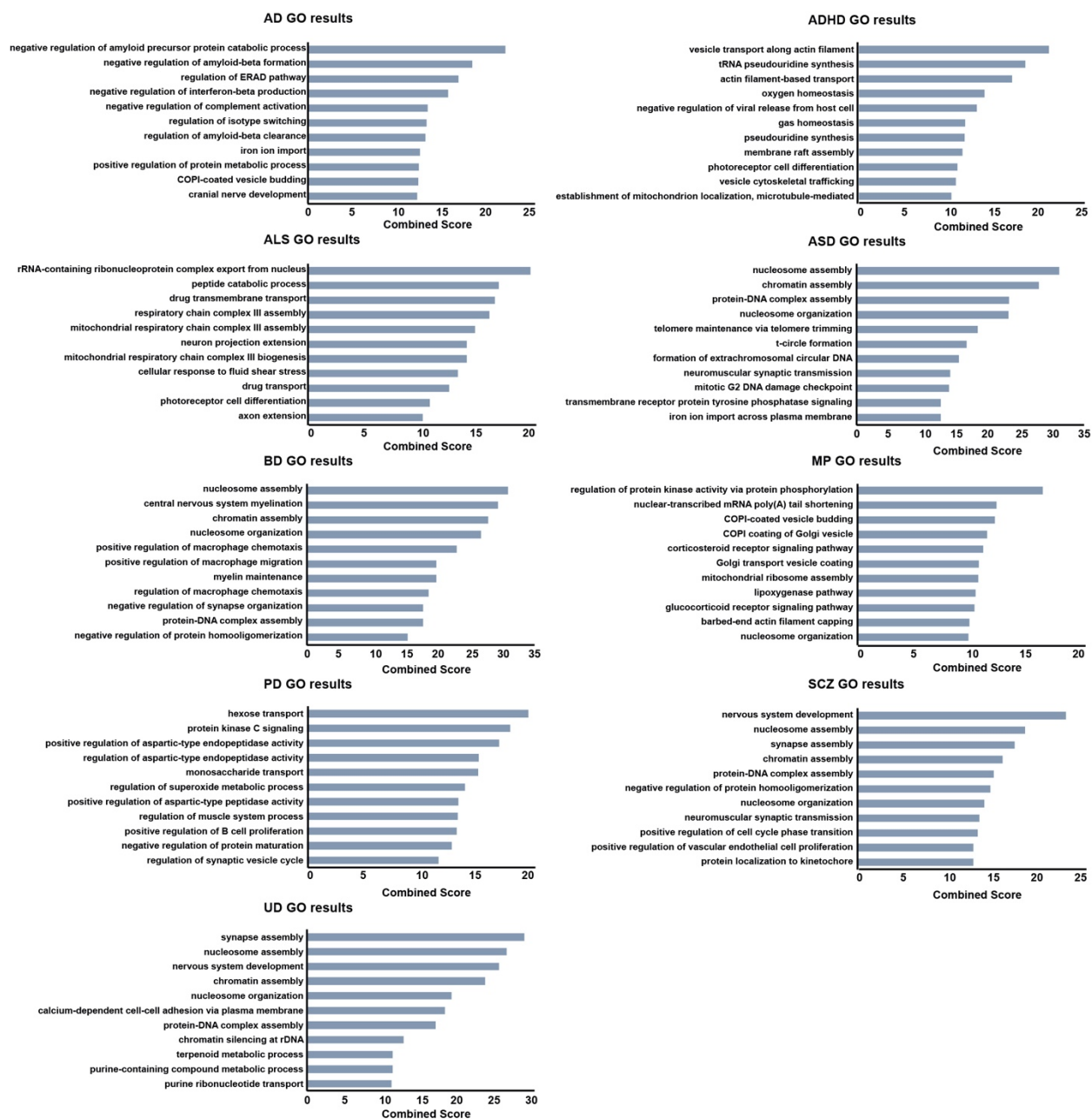
2014). Interestingly, PD SNPs are enriched in astrocytes but not in the other cell types. This enrichment of PD SNPs at astrocyte-specific PIRs supports the theory that astrocytes play an initiating role in PD, based on evidence that numerous genes implicated in PD have functions unique to astrocyte biology, as well as the neuroprotective roles astrocytes serve for dopaminergic neurons in the substantia nigra (Booth et al., 2017). EP and FTD SNPs are not enriched in any of the cell types, indicating their potential functions in alternative cell types, insufficient study power or mechanisms acting outside of chromatin-mediated gene regulation.



**Figure 1.15. Genetic analysis of chromatin interactions with complex neuropsychiatric disorder-associated variants.** (a) Enrichment analysis for complex neuropsychiatric disorder-associated variants in Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), amyotrophic lateral sclerosis (ALS), autism spectrum disorder (ASD), bipolar disorder (BP), epilepsy (EP), frontotemporal dementia (FTD), mental process (MP), Parkinson's disease (PD),

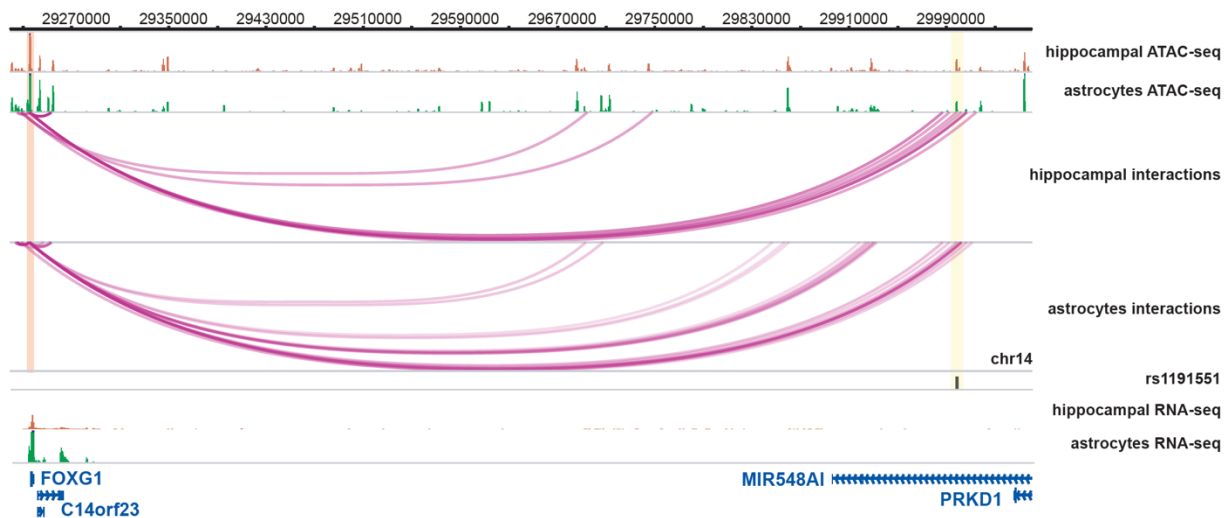
schizophrenia (SCZ), and unipolar depression (UD). The color and size of each dot represent the enrichment p-value (two-tailed one sample z-test) and the raw fold enrichment (determined as the number of SNPs overlapping significant PIRs divided by the mean number of SNPs overlapping  $n=100$  sets of randomly shuffled PIRs), respectively. **(b)** Proportions and total counts of GWAS SNPs with at least one linked SNP participating in chromatin interactions across all four cell types. **(c)** Counts of GWAS SNPs across all diseases with at least one linked SNP interacting exclusively with their nearest genes (blue), more distal genes (pink), or both (orange) and the number of target genes for each scenario (right).

Up to 70% of GWAS SNPs have at least one linked SNP overlapping PIRs in one or more cell types (**Figure 1.15b**). As it is common practice to assign GWAS SNPs to their nearest genes, we counted the number of GWAS SNPs with at least one linked SNP interacting with their nearest gene across all diseases. We find that 1,365 GWAS SNPs interact exclusively with 3,361 more distal genes (scenario I), 1,243 GWAS SNPs interact with both their nearest genes and 12,070 more distal genes (scenario II), and 248 GWAS SNPs interact exclusively with their nearest genes (scenario III) (**Figure 1.15c**). In total, 16,471 non-neighboring gene targets are identified across all diseases. To prioritize variants potentially disrupting regulatory interactions, we focused on SNPs overlapping open chromatin peaks at PIRs and find that these putative regulatory SNPs interact with genes that are relevant in the context of their respective disease etiologies. GO enrichment analysis for genes targeted by AD SNPs yields terms associated with amyloid beta formation, interferon beta production, and cranial nerve development (**Figure 1.16**). Meanwhile, genes targeted by ASD, BD, SCZ, and UD SNPs are enriched for epigenetic terms including chromatin assembly, nucleosome assembly, and nucleosome organization. For genes targeted by GWAS SNPs in the other diseases, enriched terms include neuronal processes such as myelin maintenance, neuron projection extension, synapse assembly, synaptic transmission, and nervous system development.

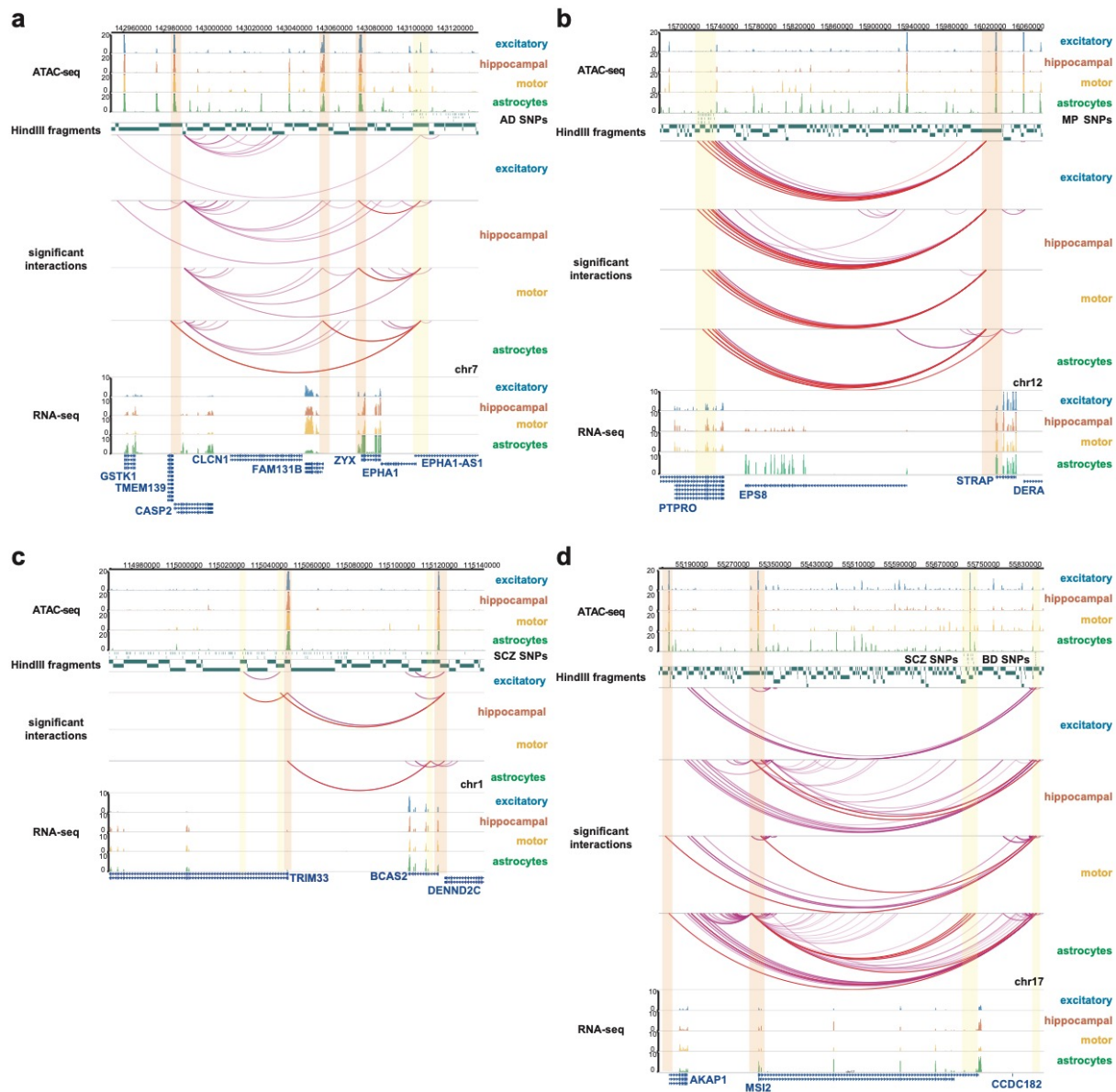


**Figure 1.16. GO enrichment analysis of PIRs with complex neuropsychiatric disorder-associated variants.** Top enriched GO terms from Enrichr for genes whose promoters are targeted by variants for each disease. EP and FTD are omitted due to their low numbers of reported variants and target genes identified by significant promoter-PIR interactions. Expanded lists of enriched GO terms are available in Song et al., 2019.

Notably, a previously reported interaction between the *FOXP1* promoter and a PIR with SCZ SNPs over 700 kb away is recapitulated by our data (Won et al., 2016) (**Figure 1.17**). At a different locus, an astrocyte-specific PIR with AD SNPs targets the promoter of *CASP2*, which encodes a well-known mediator of apoptosis that is linked to neurodegeneration (Tiwari et al., 2011; Zhao et al., 2016) (**Figure 1.18a**). Hippocampal DG-like neuron-specific PIRs with ASD SNPs target the promoter of *BCAS2*, whose knockdown in mice leads to microcephaly-like phenotypes with reduced learning, memory, and DG volume (Huang et al., 2016) (**Figure 1.18c**). Finally, the *MSI2* promoter is targeted by an astrocyte-specific PIR with SCZ SNPs, as well as by PIRs with BD SNPs in hippocampal DG-like neurons, lower motor neurons, and astrocytes (**Figure 1.18d**). Overall, we demonstrate that an approach leveraging epigenomic data to jointly prioritize and map regulatory targets for variants enables the identification of putative processes that are disrupted in disease and development.



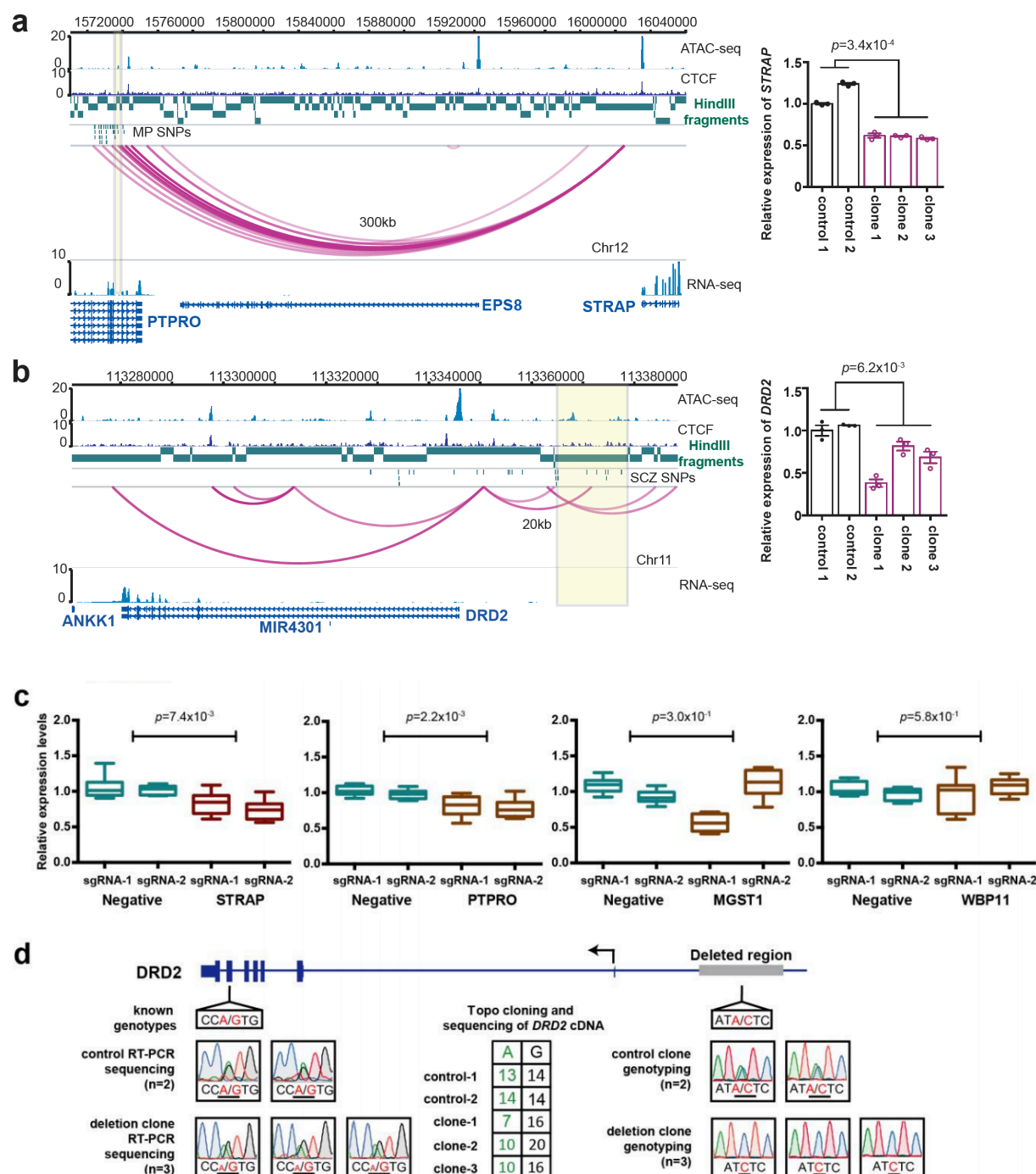
**Figure 1.17. SCZ-associated variants at the *FOXG1* locus.** Significant promoter-PIR interactions in hippocampal DG-like neurons and astrocytes recapitulate a previously reported interaction between the *FOXG1* promoter and a distal open chromatin peak containing rs1191551:G>T, a SCZ-associated variant.



**Figure 1.18. Cell type-specific PIRs with putative regulatory SNPs.** In all examples, interacting PIRs are highlighted in yellow and the targeted promoters are highlighted in orange. **(a)** A PIR with AD SNPs interacts with the promoters of *FAM131B* and *CASP2* in astrocytes, but it interacts instead with the *ZYX* promoter in hippocampal DG-like neurons and lower motor neurons. **(b)** PIRs with MP SNPs in an intron for *PTPRO* interact with the *STRAP* promoter across all four cell types. **(c)** A PIR with SCZ SNPs interacts with the *TRIM33* promoter in astrocytes. Two additional PIRs with SCZ SNPs interact with the promoters of *TRIM33* and *BCAS2* in hippocampal DG-like neurons. **(d)** A PIR with BD SNPs interacts with the *MSI2* promoter in hippocampal DG-like neurons, lower motor neurons, and astrocytes, while also interacting with the *AKAP1* promoter in lower motor neurons and astrocytes. Another group of PIRs with SCZ SNPs interacts with the *MSI2* promoter in astrocytes.

## 1.6 Validation of PIRs containing neuropsychiatric disorder risk variants

PIRs with MP SNPs in an intron for *PTPRO* interact over 300 kb away with the promoter of *STRAP* (**Figures 1.18b** and **1.19a**), which encodes a component of the survival of motor neuron (SMN) complex (Grimmler et al., 2005). The complex itself facilitates spliceosome assembly and is associated with spinal muscular atrophy (Burghes et al., 2009). To validate this locus, we derived three independent i<sup>3</sup>N iPSC clones containing biallelic deletions for a PIR in this region and observed significant downregulation of *STRAP* expression following differentiation of the cells into excitatory neurons ( $P=3.4 \times 10^{-4}$ , two-tailed two sample t-test). Targeting the same PIR with CRISPRi also consistently downregulated *STRAP* expression in excitatory neurons ( $P=7.4 \times 10^{-3}$ , two-tailed two sample t-test) (**Figure 1.19c**). Next, we focused on a PIR 20 kb upstream from the promoter of *DRD2*, which encodes the D2 subtype of the dopamine receptor. Previously, rs2514218:C>T, a noncoding variant 47 kb upstream from *DRD2*, was found to be associated with antipsychotic drug response in a cohort of schizophrenia patients (Zhang et al., 2015). This variant is in LD with a cluster of SCZ SNPs overlapping open chromatin peaks in the PIR for *DRD2*. *DRD2* is also the gene associated with the Taq1A polymorphism, which has been linked to reduced dopamine receptor density as well as addiction, anxiety, depression, and social problems in patients (Eisenstein et al., 2016). We first demonstrate that monoallelic deletion of this PIR in three independent clones leads to significant downregulation of *DRD2* expression in excitatory neurons ( $P=6.2 \times 10^{-3}$ , two tailed two sample t-test) (**Figure 1.19b**). Next, through TOPO cloning and genotyping cDNA with allele-specific variants, we confirm that monoallelic deletion of the same PIR leads to allelically imbalanced *DRD2* expression (**Figure 1.19d**). By prioritizing and validating PIRs containing putative regulatory SNPs for key genes such as *DRD2*, our approach enables the development of novel therapeutic and diagnostic strategies targeting specific risk variants in otherwise recalcitrant complex neuropsychiatric disorders.

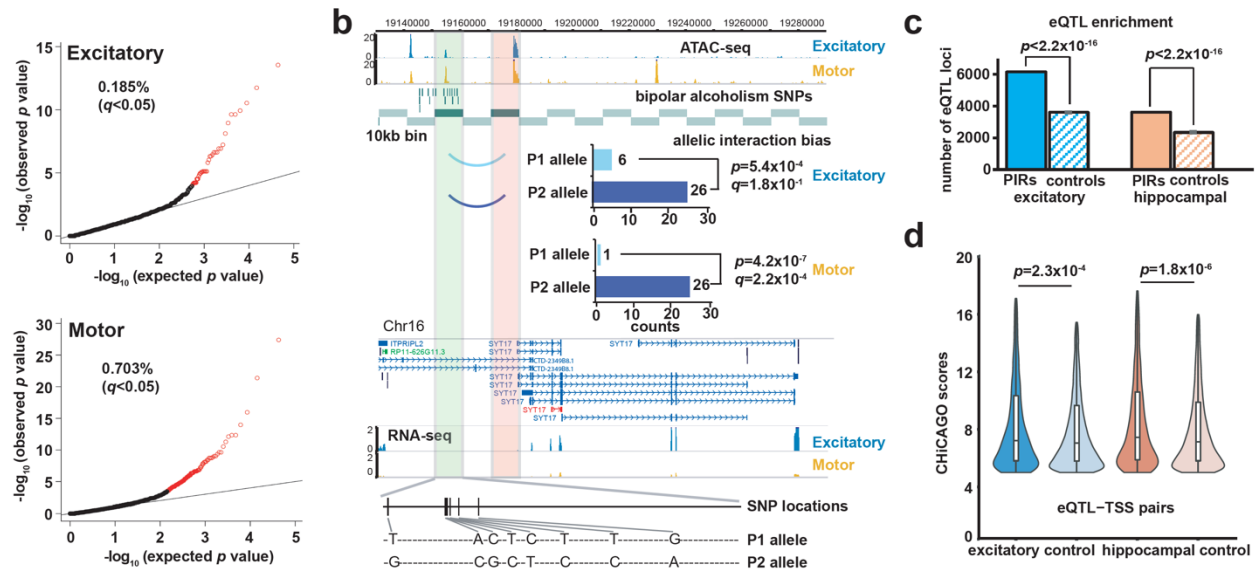


**Figure 1.19. Validation of PIRs with complex neuropsychiatric disorder-associated variants.** (a) PIRs containing MP SNPs in an intron for *PTPRO* interact with the *STRAP* promoter. Biallelic deletion of this PIR in three independent clones results in significant downregulation of *STRAP* expression in excitatory neurons ( $P=3.4 \times 10^{-4}$ ,  $t=18.5$ ,  $df=3$ , two-tailed two sample t-test). Error bars represent the SEM. (b) A PIR containing SCZ SNPs interacts with the *DRD2* promoter. Monoallelic deletion of this PIR in three independent clones results in significant downregulation

of *DRD2* expression in excitatory neurons ( $P=6.2 \times 10^{-3}$ ,  $t=6.92$ ,  $df=3$ , two tailed two sample t-test). Error bars represent the SEM. **(c)** CRISPRi silencing of a PIR for *STRAP* results in significant downregulation of *STRAP* expression in excitatory neurons ( $P=7.4 \times 10^{-3}$ ,  $t=3.34$ ,  $df=10$ , two-tailed two sample t-test). No significant downregulation was detected for the neighboring genes *MGST1* and *WBP1*, though the expression of *PTPRO* was affected ( $P=3.0 \times 10^{-1}$ ,  $t=1.09$ ,  $df=10$ ,  $P=5.8 \times 10^{-1}$ ,  $t=0.570$ ,  $df=10$ , and  $P=2.2 \times 10^{-3}$ ,  $t=4.07$ ,  $df=10$ , respectively, two-tailed two sample t-test). Three independent replicates per condition and two sgRNAs per replicate were used for each experiment. Boxplots indicate the median, IQR, minimum, and maximum. **(d)** Schematic of detected genotypes in the *DRD2* gene and its PIR in two wild type clones and three monoallelic deletion clones. Genotyping and qPCR sequencing for WTC11 variants in the *DRD2* gene reveal allele-specific imbalances in *DRD2* expression, consistent with the monoallelic deletion of the PIR.

## 1.7 Genetic variants contribute to chromatin interaction bias and changes in transcription

Since regulatory variants and other genetic perturbations are thought to introduce or disrupt chromatin loops between promoters and PIRs, we were interested to see if we could detect instances of allelic bias across our sets of significant promoter-PIR interactions. We used our chromatin interaction data to perform genome-wide phasing of WTC11 variants using HaploSeq (Selvaraj et al., 2013) and performed allele-specific mapping at a resolution of 10 kb using HiC-Pro (Servant et al., 2015). We identified 41 (0.185%) and 151 (0.703%) of significantly interacting bins to exhibit allelic bias at an FDR cutoff of 5% (two-tailed binomial test with BH correction) in the excitatory and lower motor neurons, respectively (**Figure 1.20a**). In one case, allelically biased interactions were detected between a PIR containing bipolar alcoholism SNPs (Lydall et al., 2011) and the promoter of *SYT17*, which encodes a member of a family of membrane-trafficking proteins that mediate synaptic function and calcium-controlled neurotransmitter release (Sudhof et al., 2012). The risk allele of the lead variant (rs8062326:G>A) is associated with the WTC11 allele exhibiting reduced interaction frequency in both cell types (**Figure 1.20b**), suggesting that regulatory variants can increase individual risk for bipolar alcoholism by disrupting interactions for *SYT17*.



**Figure 1.20. Genetics variants contribute to chromatin interaction bias and alterations in gene expression.** (a) Quantile-quantile plots showing the proportions of interacting 10 kb bins exhibiting significant allelic bias at an FDR cutoff of 5% (two-tailed binomial test with BH correction) in excitatory neurons (n=22,162 bins) and lower motor neurons (n=21,479 bins). (b) A sample interaction with significant allelic bias in excitatory neurons ( $P=5.4 \times 10^{-4}$ , two-tailed binomial test) and lower motor neurons ( $P=4.2 \times 10^{-7}$ , two-tailed binomial test). The interaction occurs between the SYT17 promoter and a PIR with bipolar alcoholism SNPs at an open chromatin peak. Heterozygous phased WTC11 variants at the PIR as well as bar graphs of allele-specific read counts are shown. (c) Enrichment of significant eQTLs from GTEx V7 at significant versus randomly shuffled PIRs in matched tissue types for excitatory and hippocampal DG-like neurons ( $P < 2.2 \times 10^{-16}$  for both cell types, two-tailed one sample z-test). Means and the SEM for the number of overlaps across n=100 sets of randomly shuffled PIRs are shown. (d) Distributions of interaction scores for chromatin interactions overlapping significant eQTL-TSS pairs versus randomly sampled nonsignificant eQTL-TSS pairs in excitatory neurons and hippocampal DG-like neurons ( $P=2.3 \times 10^{-4}$  for excitatory neurons and  $P=1.8 \times 10^{-6}$  for hippocampal DG-like neurons, two-tailed two sample Kolmogorov-Smirnov test). Violin plots show the distributions of gene expression values within each group, and boxplots indicate the median, IQR, Q1-1.5\*IQR, and Q3+1.5\*IQR.

Physical chromatin interactions have been theorized to mediate the effects of cis-acting regulatory variants including expression quantitative trait loci (eQTLs) on gene expression. In support of this hypothesis, we first demonstrate that significant eQTLs in cortical and hippocampal tissues from GTEx V7 (GTEx Consortium, 2017) are enriched at PIRs for excitatory and hippocampal DG-like neurons, respectively ( $P < 2.2 \times 10^{-16}$  for both cell types, two-tailed one sample z-test) (**Figure**

**1.20c).** Next, we show that scores for interactions overlapping significant eQTL-TSS pairs are significantly higher than scores for interactions overlapping randomly shuffled eQTL-TSS pairs ( $P=2.28 \times 10^{-4}$  for excitatory neurons and  $P=1.76 \times 10^{-6}$  for hippocampal DG-like neurons, two-tailed two sample Kolmogorov-Smirnov test) (**Figure 1.20d**). This indicates that chromatin interactions recapitulating regulatory relationships between significant eQTL-TSS pairs are identified by pcHi-C with increased levels of confidence. Overall, our results present orthogonal lines of evidence that chromatin interactions can not only be altered by variants in an allele-specific manner, but that variants can also modulate gene expression through the formation or disruption of regulatory chromatin loops.

## 1.8 Discussion

Here, we have leveraged pcHi-C, ATAC-seq, and RNA-seq to comprehensively annotate previously uncharted regulatory relationships between promoters and distal regulatory elements in cell types that are relevant to complex neuropsychiatric disorders. We show that PIRs are not only cell type-specific, but they are enriched for regulatory chromatin signatures including open chromatin peaks and in vivo-validated enhancer elements from the VISTA Enhancer Browser. Inspection of cell type-specific distal open chromatin peaks at PIRs reveals subtype-specific binding sites for TFs involved in the specification and maintenance of cellular identity. Furthermore, our interactions identify novel gene targets for disease-associated variants and enable the prioritization of variants for validation using CRISPR techniques. We report a large number of putative regulatory variants which may provide additional insights into aspects of complex disease biology. Finally, the disease- and cell type-specific enrichment of variants at PIRs, combined with the observation that the same PIRs can target different genes in different cell types, supports existing evidence that regulatory variants possess context-dependent functional specificities.

The integrative analysis in this study has several limitations including a lack of cell type-specific annotations for genomic and epigenomic features occurring at PIRs. For example, the analysis of chromatin state and eQTL enrichment at PIRs used data in matched tissues from the Roadmap Epigenomics Project and GTEx V7, respectively. Furthermore, while studying chromatin interactions in healthy cells enables the detection of regulatory interactions in the absence of dysregulation, the epigenomic characterization of patient-derived cells will be important to glean specific insights into how the 3D epigenome is altered in disease. Additional experiments are necessary to determine how the haploinsufficiency of genes such as *STRAP* and *DRD2* may contribute to phenotypes in disease. Finally, in vitro cultured cells can at present only approximate the full set of cellular responses occurring in vivo, especially for complex structures such as the brain, and they may reflect different developmental stages than expected based on their time in culture. Future approaches isolating specific cell types from tissues, using single cell sequencing, or using advanced organoid models will be essential for drilling down more deeply into mechanisms driving cellular identity and disease. The epigenomic characterization of additional cell types should continue to yield rich insights into the landscape of transcriptional regulation, contributing towards an improved understanding of complex disease biology.

## **1.9 Methods**

### **Cell culture**

Human excitatory neurons were generated using integrated, isogenic, and inducible neurogenin-2 (*Ngn2*) iPSCs (i<sup>3</sup>N iPSCs) with doxycycline-inducible mouse *Ngn2* integrated at the AAVS1 safe-harbor locus. The i<sup>3</sup>N iPSCs have a well-characterized wild type genetic background (WTC11) (Miyaoaka et al., 2014). A simplified, two-step pre-differentiation and maturation protocol was used to generate the excitatory neurons (Wang et al., 2017). Briefly, i<sup>3</sup>N iPSCs were

incubated with 2 µg/ml doxycycline in pre-differentiation media containing KnockOut DMEM/F12 supplemented with 1x N-2, 1x NEAA, 1 µg/ml mouse laminin, 10 ng/ml BDNF, and 10 ng/ml NT3. In addition, 10 µM Rock inhibitor was included in the pre-differentiation media for the first day. Media was changed daily for three days. For maturation, pre-differentiated precursor cells were dissociated and subplated on poly-D-lysine/laminin plates in maturation media containing equal parts DMEM/F12 and Neurobasal-A with 2 µg/ml doxycycline and supplemented with 0.5x B-27, 0.5x N-2, 1x NEAA, 0.5x GlutaMax, 1 µg/ml mouse laminin, 10 ng/ml BDNF, and 10 ng/ml NT3. The doxycycline was omitted from all subsequent media changes. Half of the media was changed weekly for the first two weeks, then the amount of media was doubled on day 21. Thereafter, a third of the media was replaced weekly until harvesting. 7 to 8 week old excitatory neurons were harvested for library preparation.

Human hippocampal DG-like neurons were generated from dissociated hippocampal organoids. Briefly, WTC11 iPSCs were grown on MEF feeder cells and patterned towards a neural ectoderm fate using dual SMAD inhibition as floating embryoid bodies (EBs) in media containing 20% KnockOut Serum Replacement. Next, 4 week old EBs were patterned towards a hippocampal fate using WNT and BMP in media containing 1x N-2. After patterning, organoids were dissociated using a neural tissue dissociation kit (MiltenyiBiotec), plated on PDL- and laminin-coated plates, and cultured for 4 weeks in media containing 1x B-27, 10 ng/ml BDNF, 10 ng/ml GDNF, 0.5 mM cAMP, and 200 µM ascorbic acid.

Human lower motor neurons were differentiated from WTC11 iPSCs with a doxycycline inducible transgene expressing *NGN2*, *ISL1*, and *LHX3* integrated at the AAVS1 safe-harbor locus (i<sup>3</sup>LMN iPSCs) (Fernandopulle et al., 2018). Briefly, i<sup>3</sup>LMN iPSCs were maintained on growth factor reduced Matrigel in StemFit media (Nacalai USA). On day 0, 1.5\*10<sup>6</sup> i<sup>3</sup>LMN iPSCs were plated on 10-cm dishes, followed 24 hours later by exchange into neural induction media containing

doxycycline and compound E. On day 3, the precursor cells were re-plated onto 12-well plates coated with poly-D-lysine and laminin at a density of  $2.5 \times 10^5$  cells per well. From day 3 to day 4, the cells were treated with a pulse of 40  $\mu$ M BrdU for 24 hours to suppress the proliferation of undifferentiated cells. Media was exchanged on day 4 and every three days thereafter. The cells were harvested after 10 days post-differentiation for library preparation.

Human primary astrocytes (P0) were purchased from ScienCell Research Laboratories (catalog #1800) and cultured using the recommended media (catalog #1801). Briefly, cells were cultured in flasks coated with poly-L-lysine (2  $\mu$ g/cm<sup>2</sup>) and passaged once with trypsin and EDTA before harvesting.

### **Promoter capture Hi-C (pcHi-C)**

In situ Hi-C libraries for the excitatory neurons, hippocampal DG-like neurons, lower motor neurons, and astrocytes were constructed from 1 to 2 million cells (fixed in 1% PFA) using HindIII as a restriction enzyme as previously described (Rao et al., 2014). pcHi-C was performed using biotinylated RNA probes according to an established protocol (Jung et al., 2019). Briefly, sets of 120 bp probes with 30 bp overhangs were designed to capture sequences adjacent to the restriction sites flanking each promoter-containing HindIII fragment. Three probes were targeted to each side of a restriction site for a total of 12 probes targeting each promoter-containing HindIII fragment. In total, promoters (defined as the sequences up to 500 bp upstream and downstream of each TSS) for 19,603 of the 20,332 protein coding genes in GENCODE 19 were captured using this approach. While noncoding RNA promoters were not explicitly targeted by this design, the probes also targeted HindIII fragments containing 3,311 of the 14,069 noncoding RNA promoters in GENCODE 19.

To perform the hybridization, 500 ng of each situ Hi-C library was first mixed with 2.5 µg human Cot-1 DNA (Invitrogen #15279011), 2.5 µg salmon sperm DNA (Invitrogen #15632011), and 0.5 nmol each of the p5 and p7 IDT xGen Universal Blocking Oligos in a total volume of 10 µL and denatured for 5 min at 95°C before holding at 65°C. Next, a hybridization buffer mix was prepared by combining 25 µL 20x SSPE, 1 µL 0.5 M EDTA, 10 µL 50x Denhardt's solution, and 13 µL 1% SDS and pre-warming the mix to 65°C. Finally, 500 ng of the probes was mixed with 1 µL 20 U/µL SUPERase-In (Invitrogen #AM2696) in a total volume of 6 µL, pre-warmed to 65°C, then combined with the library and hybridization buffer mixes. The solution was transferred to a humidified hybridization chamber and incubated for 24 hours at 65°C. 0.5 mg Dynabeads MyOne Streptavidin T1 magnetic beads (Invitrogen #65601) were used to pull down the captured fragments in a binding buffer consisting of 10 mM Tris-HCl pH 7.5, 1 M NaCl, and 1 mM EDTA. The beads were washed once with 1x SSC and 0.1% SDS for 30 minutes at 25°C, followed by three washes with pre-warmed 0.1X SSC and 0.1% SDS for 10 minutes at 65°C. The library was eluted in 20 µL nuclease-free water, amplified, then sent for paired-end sequencing on the HiSeq 4000 (50 bp reads), the HiSeq X Ten (150 bp reads), or the NovaSeq 6000 (150 bp reads).

### **Calling significant promoter-PIR interactions**

Paired-end sequencing reads were first trimmed using fastp 0.20.0 (Chen et al., 2018) running the default settings before being mapped, filtered, and deduplicated using HiCUP 0.71 (Wingett et al., 2015) with bowtie2 (Langmead et al., 2012) and filtering for ditags between 100 and 1200 bp. In addition, the sequencing depths of all libraries was normalized so that each replicate had the same number of usable reads (defined as the number of on-target cis pairs interacting over a distance of 10 kb). Significant promoter-PIR interactions were called using CHiCAGO (Cairns et al., 2016) running the default settings and retaining baited fragments that are supported by at least 250 reads (minNPerBaits=250). Interactions between HindIII fragments with a score

(defined as the negative log p-value)  $\geq 5$  in each cell type were determined to be significant. In cases where CHiCAGO reported the same interaction in different orientations, the two interactions were merged, retaining the higher score of the two interactions. Interchromosomal interactions were omitted from the analysis. To call overlaps between our sets of significant interactions and genomic and epigenomic features including promoters, open chromatin peaks, chromatin states, disease-associated variants, and eQTLs, interacting bins were expanded to a minimum width of 5 kb or retained as the original widths of the HindIII fragments if they exceeded 5 kb. Interactions involving HindIII fragments larger than 100 kb were omitted from our analysis. An interaction was considered to be shared between cell types if both of its interacting ends intersected the corresponding ends of an interaction in another cell type. Otherwise, an interaction was determined to be cell type-specific.

### **Validation of PIRs using CRISPR deletion**

To validate genomic interactions captured by pcHi-C, candidate PIRs were targeted for CRISPR deletion in the i<sup>3</sup>N iPSCs. Pairs of sgRNAs targeting the putative regulatory element as localized by open chromatin peaks in the candidate PIR were designed for each locus of interest. All sgRNAs were synthesized by Synthego. Cas9 protein was sourced from QB3-Berkeley. To generate deletion lines, CRISPR/Cas9 nucleofections were performed using the LONZA Human Stem Cell Nucleofector® Kit. For each nucleofection, approximately 500,000 i<sup>3</sup>N iPSCs were transfected with Cas9:sgRNA RNP complex (consisting of 12 µg Cas9, 10 µg sgRNA 1, and 10 µg sgRNA 2) using program “A-023” on the LONZA 4D-Nucleofector. The nucleofected cells were then seeded onto Matrigel-coated 6-well plates containing Essential 8™ Medium (ThermoFisher #A15169-01) with Y-27632 added for recovery following nucleofection. After 48 hours, the cells were split into new 6-well plates at a concentration of approximately 50 cells per well for picking single colonies. Clones picked from the 6-well plates containing homozygous deletions were

confirmed by qPCR and induced into excitatory neurons for quantifying the expression of genes targeted by the deleted PIRs. For each experiment, we used three deletion clones and two wild type clones. Total RNA from the excitatory neurons was extracted using a Qiagen AllPrep DNA/RNA Mini Kit, and cDNA was synthesized using a Bio-RAD iScript™ cDNA Synthesis Kit. qPCR for targeted genes was performed with FastStart Essential DNA Green Master reaction mix (Roche) on the LightCycler® 96 System (Roche).

### **Validation of PIRs using CRISPRi**

Excitatory neurons induced from i<sup>3</sup>N iPSCs were infected with lentivirus carrying dCas9-KRAB-blast (Addgene #89567) and colonies with high expression of dCas9 were picked. The CROP-seq-opti vector (Addgene #106280) was used for sgRNA expression. sgRNAs were cotransfected with lentivirus packaging plasmids pMD2.G (Addgene #12259) and psPAX (Addgene #12260) into 293T cells with PolyJet (SigmaGen Laboratories #SL100688) according to the manufacturer's instructions. Virus-containing media was collected for 72 hours, filtered through a 0.45 µm filter (Millipore #SLHV033RS), then concentrated using an Amicon Ultra centrifugal filter (Millipore #UFC801024). The virus was titrated in the excitatory neurons by qPCR 72 hours post-transfection. The internal qPCR control targeted an intronic region (forward primer: TCCTCCGGAGTTATTCTTGGCA and reverse primer: CCCCCATCTGATCTGTTTCAC). Integration of the WPRE fragment was quantified in comparison with a cell line containing a known copy number of WPRE. For CRISPRi silencing of putative regulatory elements, excitatory neurons were treated with lentivirus containing the sgRNAs in two or three independent differentiations (MOI ~3). Cells were collected for mRNA extraction 7 days post-transfection, and gene expression was determined using qPCR. Each independent replicate was repeated with three technical replicates, and the mean values were used for statistical testing.

## **Motif enrichment analysis**

We took the sets of all cell type-specific distal open chromatin peaks participating in significant promoter-PIR interactions between promoter-containing and non-promoter-containing bins for each cell type and used the sequences in 250 bp windows around the peak summits to perform motif enrichment analysis using HOMER (Heinz et al., 2010) running the default settings. The cumulative binomial distribution was used for motif scoring. The entire genome was used as a background. Significance and expression values for each detected motif and its corresponding TFs are reported in Song et al., 2019.

## **VISTA enhancer analysis and target gene identification**

Human and mouse enhancer regions with orthologous human sequences and positive annotations in the VISTA Enhancer Browser (Visel et al., 2007) were downloaded and analyzed for overlap with our sets of significant promoter-PIR interactions for each cell type. Of the 2,956 tested elements in their database (January 2019), 1,568 were positive elements with orthologous human sequences (976 human elements and 892 mouse elements with orthologous human sequences). Positive elements were expanded to a minimum width of 5 kb. The “hs” or “mm” prefixes for elements indicate the species of DNA origin. For determining whether positive elements interacted with their nearest or more distal genes, we only considered protein coding and noncoding RNA genes in GENCODE 19. To evaluate cases where interactions between positive elements and their nearest genes were unresolvable (“same fragment ambiguity”), we determined if a promoter for the nearest gene overlapped at least one HindIII fragment that the positive element did not also overlap. The following terms were considered to be neural annotations: neural tube, hindbrain, cranial nerve, midbrain, forebrain, mesenchyme derived from neural crest, dorsal root ganglion, and trigeminal V.

## **SNP enrichment analysis and target gene identification**

GWAS SNPs for a total of eleven neuropsychiatric disorders including Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), amyotrophic lateral sclerosis (ALS), autism spectrum disorder (ASD), bipolar disorder (BD), epilepsy (EP), frontotemporal dementia (FTD), mental process (MP), Parkinson's disease (PD), and schizophrenia (SCZ), and unipolar depression (UD) were mined from the GWAS Catalog (Buniello et al., 2019) (December 2018) using a p-value threshold of  $10^{-6}$ . The GWAS SNPs were expanded to sets of linked SNPs using HaploReg 4.140 at an LD threshold of 0.8 according to the reported study population(s) for each SNP. All SNPs were lifted over to hg19 and filtered for duplicates by position. Disease- and cell type-specific enrichment for SNPs was calculated as the number of SNPs overlapping significant PIRs divided by the mean number of SNPs overlapping randomly shuffled PIRs with matching distance distributions.  $n=100$  sets of randomly shuffled PIRs were sampled in each case. To determine whether a GWAS SNP interacted with a target gene, we determined whether it or any of its linked SNPs (expanded to a minimum width of 1 kb) interacted with a promoter associated with the nearest gene. To evaluate cases where interactions between GWAS SNPs and their nearest genes were unresolvable ("same fragment ambiguity"), we determined if a promoter for the nearest gene overlapped at least one HindIII fragment that a GWAS SNP or any of its linked SNPs did not also overlap. Finally, we derived a list of SNPs for which the SNP was located within 2 kb of the center of an open chromatin peak at a PIR, indicating additional evidence for a functional regulatory variant at that locus. These SNPs are referred to as "putative regulatory SNPs."

## **Phasing of the WTC11 genome**

Phasing of the WTC11 genome was performed as previously described (Selvaraj et al., 2013). Briefly, WTC11 variants were split by chromosome and phase-informative reads from pChI-C were extracted using extractHAIRS with the minimum mapping quality set to 10 and the maximum insert size set to 30000000 bp (Bansal et al., 2008). Phasing was performed with Hapcut using a maximum of 101 iterations. Next, we extracted the maximum variants phased (MVP) haplotype block from the output of Hapcut to use as a seed haplotype. We modified the “neighborhood correction” aspect of phasing by filtering phased variants whose predicted phase would have a marginal probability below 0.99 according to an in-house implementation of a hidden Markov model (HMM) as described previously (Li et al., 2003; Delaneau et al., 2013) with a reference haplotype set from the 1000 Genomes Project. Missing variants were imputed using the aforementioned HMM with the reference haplotype set from the 1000 Genomes Project. The WTC11 SNP phasing data is available at the Gene Expression Omnibus under the following accessible number: GSE113483.

## **Allelic bias analysis**

We used the WTC11 phasing data along with the allele-specific mapping capabilities of HiC-Pro (Servant et al., 2015) to quantify allelic bias between significantly interacting 10 kb bins genome-wide in the excitatory and lower motor neurons. We selected these two cell types because they used homogenous induction of TFs for differentiation, therefore minimizing the noise introduced by conventional differentiation techniques. Briefly, reads were mapped using bowtie (Langmead et al., 2009) to a version of the hg19 reference genome where all sites with heterozygous phased SNPs were masked. Unfiltered HiC-Pro contact maps were used for this analysis. Next, nucleotides at masked polymorphic sites were used to assign reads to either allele. Reads

reporting conflicting allele assignments or unexpected bases were filtered out. Reads with at least one allele-specific mate were used to construct allele-specific Hi-C contact maps at a resolution of 10 kb. The allele-specific Hi-C contact maps were intersected with the set of all significant promoter-PIR interactions with score  $\geq 3$  to assess allelic bias between interacting 10 kb bins. Only interacting bins with 10 or more reads across both alleles were retained (n=22,162 bins for excitatory neurons and n=21,479 bins for lower motor neurons). A two-tailed binomial test was used to assess allelic bias across each set of interacting bins, and the resulting p-values were adjusted using BH correction to filter out significantly biased loci at an FDR cutoff of 5%. Allelically biased interactions with p-values  $< 10^{-3}$  are reported in Song et al., 2019.

### **eQTL enrichment analysis**

1D enrichment of significant eQTLs from GTEX V7 (GTEx Consortium, 2017) at significant versus randomly shuffled PIRs in matched tissue types for excitatory neurons (Brain - Cortex, n=136) and hippocampal DG-like neurons (Brain - Hippocampus, n=111) was performed in the same manner as the chromatin state and SNP enrichment analysis. We used the full set of significant eQTL SNP-gene associations available from GTEx which included 478,903 eQTLs associated with 6,146 eGenes for the cortex and 221,876 eQTLs associated with 3,262 eGenes for the hippocampus. To determine the 2D enrichment of eQTL-TSS pairs in our significant interaction sets, we first filtered out eQTL-TSS pairs that were within 10 kb of each other or on the same HindIII fragment as this would be below the minimum detectable resolution by pcHi-C. Next, we sampled a set of nonsignificant eQTL-TSS pairs with a matching distance distribution as the set of significant eQTL-TSS pairs for each cell type, controlling for the number of genes around which the eQTL-TSS pairs were centered. We compared the distributions of interaction scores for significant interactions supporting the significant and nonsignificant eQTL-TSS pairs by overlapping the eQTL-TSS pairs with our significant interactions.

## **Chapter 2. Cell Type-Specific 3D Epigenomes in the Developing Human Cortex**

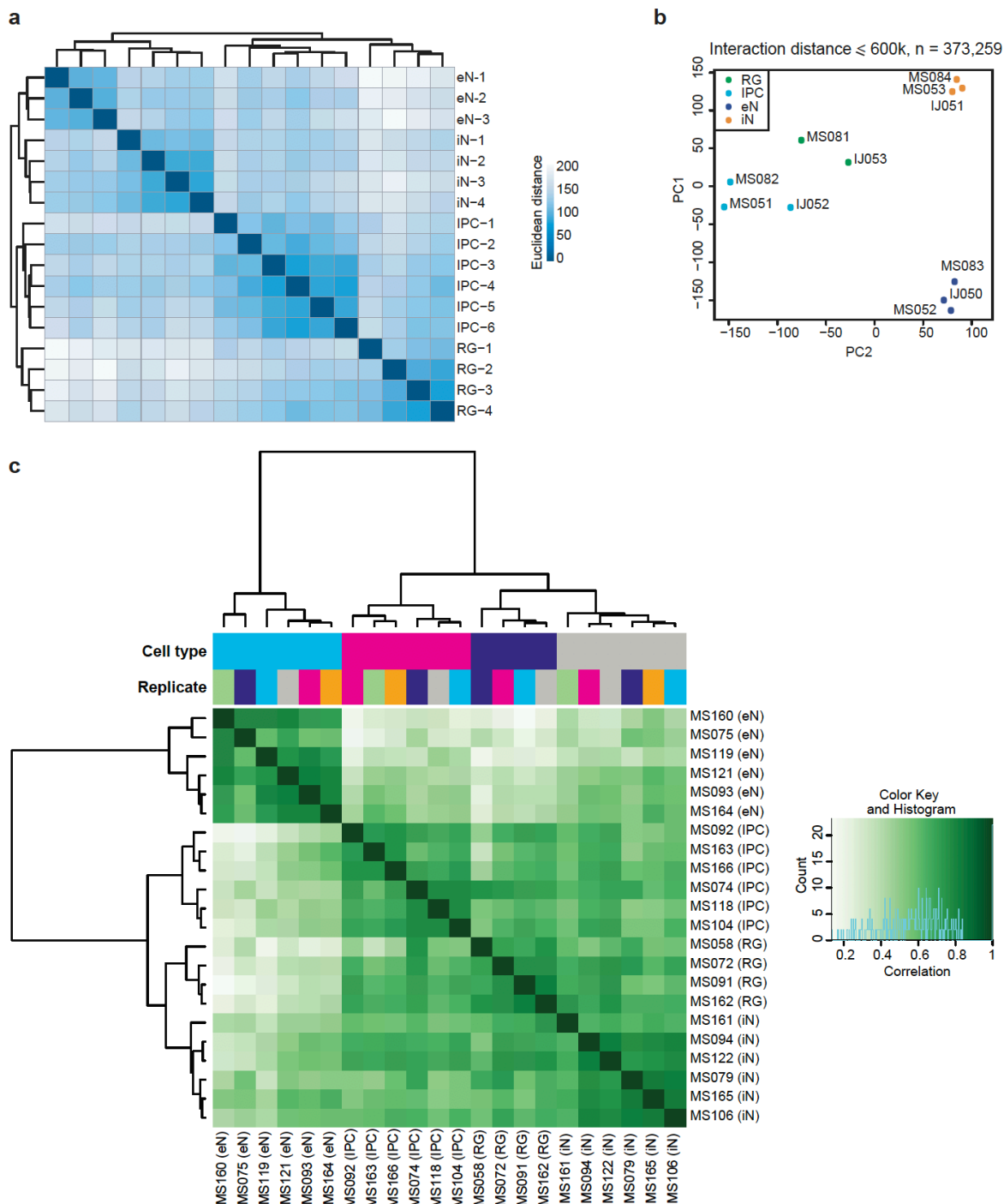
The human cortex is a complex, heterogeneous structure which undergoes extensive expansion during development, a process which is markedly different and features distinct cell types from mouse cortex development. Previous studies utilized single cell RNA sequencing (scRNA-seq) to unravel the transcriptomic diversity of the developing human cortex, revealing at least nine major cell types and up to 26 distinct subtypes in the dorsal cortex alone (Nowakowski et al., 2017; Zhong et al., 2018). Much of this diversity arises from cortical stem cells known as radial glia (RG), whose cell bodies reside in the germinal zones (GZs) of the dorsal and ventral cortices. Within the dorsal cortex, RG divide asymmetrically to give rise to intermediate progenitor cells (IPCs), which proliferate and differentiate into excitatory neurons (eNs) (Pontious et al., 2008; Hansen et al., 2010). These newborn neurons undergo radial migration until they reach the cortical plate (CP), where they mature and undergo synaptogenesis. Meanwhile, interneurons (iNs) produced in the ventral cortex migrate tangentially into the dorsal cortex through the marginal and germinal zones (Anderson et al., 1999). These processes result in a CP comprised primarily of eNs and iNs, and a GZ where all four cell types are intermixed.

Dynamic changes in the epigenomic landscape have been shown to play a critical role in development and cell fate commitment, for instance through the rewiring of physical chromatin loops between promoters and distal regulatory elements (Zheng et al., 2019). These regulatory interactions are of particular interest as their dysregulation has been linked to complex disorders and traits (Li et al., 2018; Schoenfelder et al., 2019). Despite their utility, detailed epigenomic characterizations are still missing for specific cell types in the developing human cortex due to limitations associated with the analysis of bulk tissues (Won et al., 2016). For example, previous studies focused on the CP and GZ, both of which contain diverse mixtures of lineages encompassing endothelial cells, microglia, and oligodendrocyte progenitor cells in addition to RG,

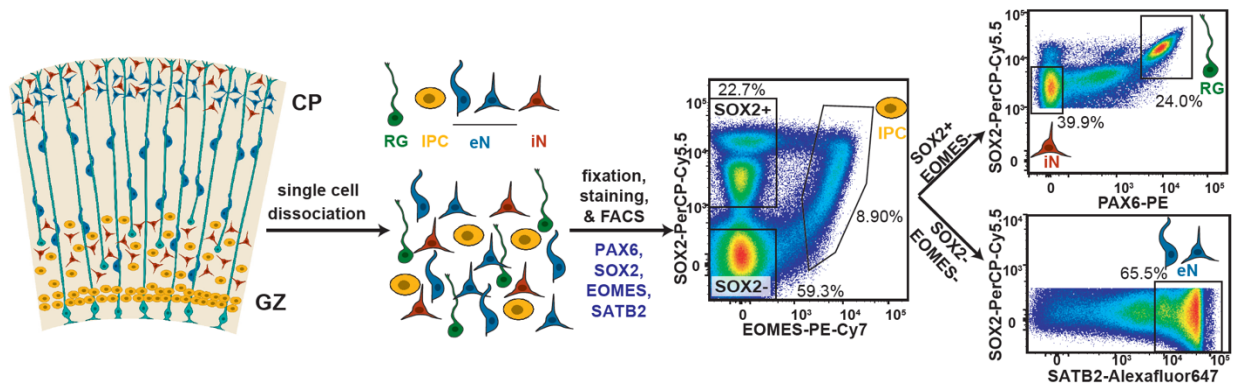
IPCs, eNs, and iNs. Furthermore, important developmental genes such as *SOX2*, *PAX6*, and *NPAS3* are often broadly expressed across multiple cell types and lineages, obfuscating interpretations of heterogeneous data. Here, we present a novel approach for isolating RG, IPCs, eNs, and iNs from mid-gestational human cortex samples, enabling the profiling of their cell type-specific epigenomic landscapes. In addition, we present CRISPRview, a versatile and sensitive technique for validating cell type-specific cis-regulatory elements in heterogeneous populations of primary cells. We apply CRISPRview to demonstrate that key genes including *GPX3*, *TNC*, *HES1*, and *IDH1* are regulated by cell type-specific enhancers in RG and eNs. Overall, our results identify novel mechanisms underlying gene regulation and lineage specification during human corticogenesis, providing a framework for the understanding of diverse processes in development and disease.

## **2.1 Sorting specific cell types from the developing human cortex**

To isolate RG, IPCs, eNs, and iNs from mid-gestational human cortex samples between gestational weeks (GW) 15 to 22, we expanded upon an established approach for isolating RG from human cortical samples using fluorescence-activated cell sorting (FACS) (Thomsen et al., 2016). Microdissected GZ and CP samples were dissociated, stained using antibodies for EOMES, *SOX2*, *PAX6*, and *SATB2*, and partitioned into their constituent populations using FACS (**Figure 2.2**). IPCs were isolated as the EOMES<sup>+</sup> population, while eNs were isolated from the EOMES<sup>-</sup> and *SOX2*<sup>-</sup> population based on high *SATB2* expression, which marks both newborn and mature eNs at the samples' ages (Nowakowski et al., 2017). RG were isolated based on high *SOX2* and high *PAX6* expression, and iNs were isolated based on medium *SOX2* and low *PAX6* expression. The gene expression profiles of the sorted cell populations were both highly consistent with cellular identity and reproducible between individuals (**Figures 2.1a** and **2.3a-b**).



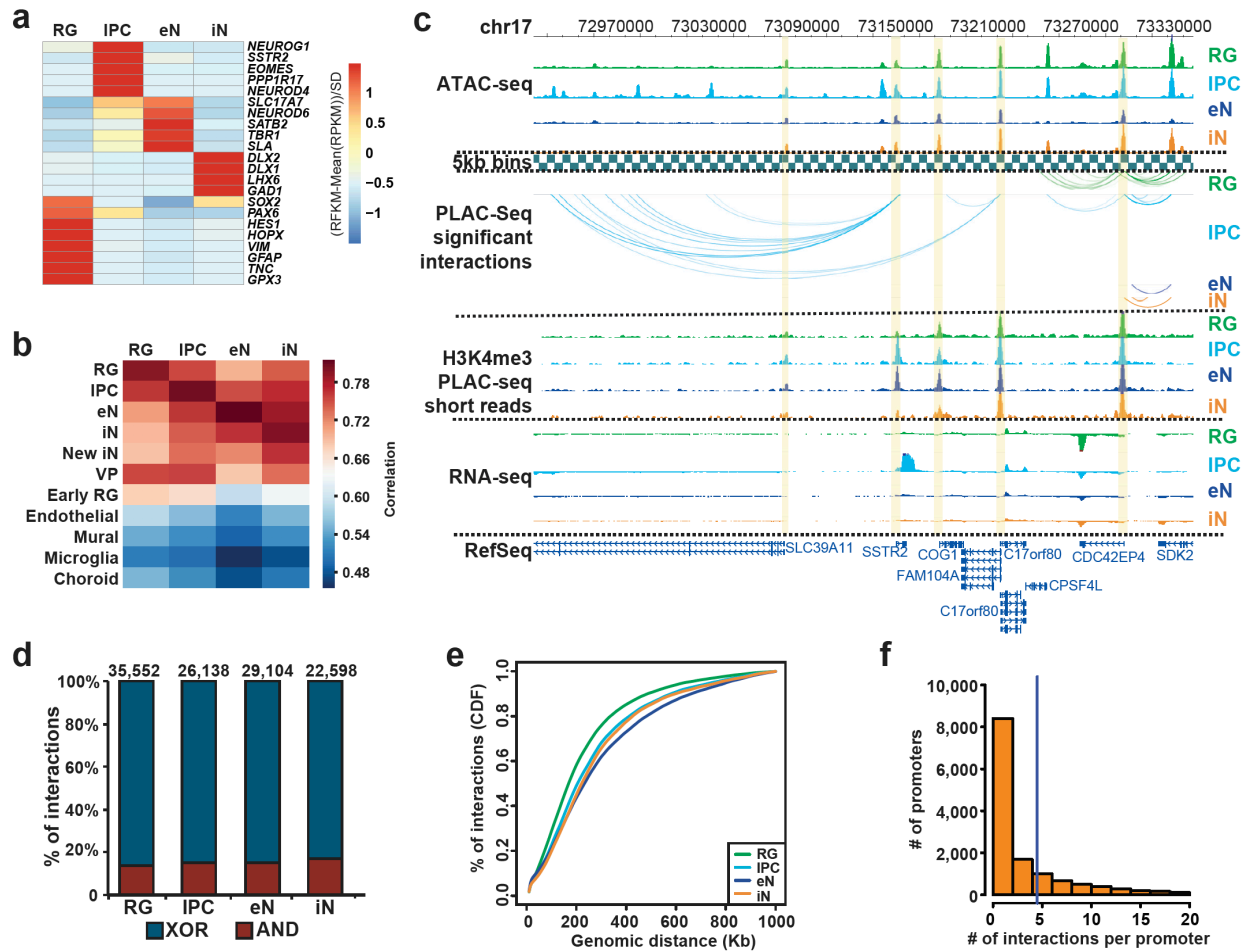
**Figure 2.1. Reproducibility between RNA-seq, ATAC-seq, and PLAC-seq replicates. (a)** RNA-seq replicates were hierarchically clustered according to gene expression sample distances using DESeq2. **(b)** Principle component analysis (PCA) was performed based on normalized contact frequencies across all PLAC-seq replicates (see methods). PCA was performed using interacting 5 kb bins in a 600 kb window. **(c)** Heatmap showing correlations and hierarchical clustering for read densities at open chromatin peaks across all ATAC-seq replicates.



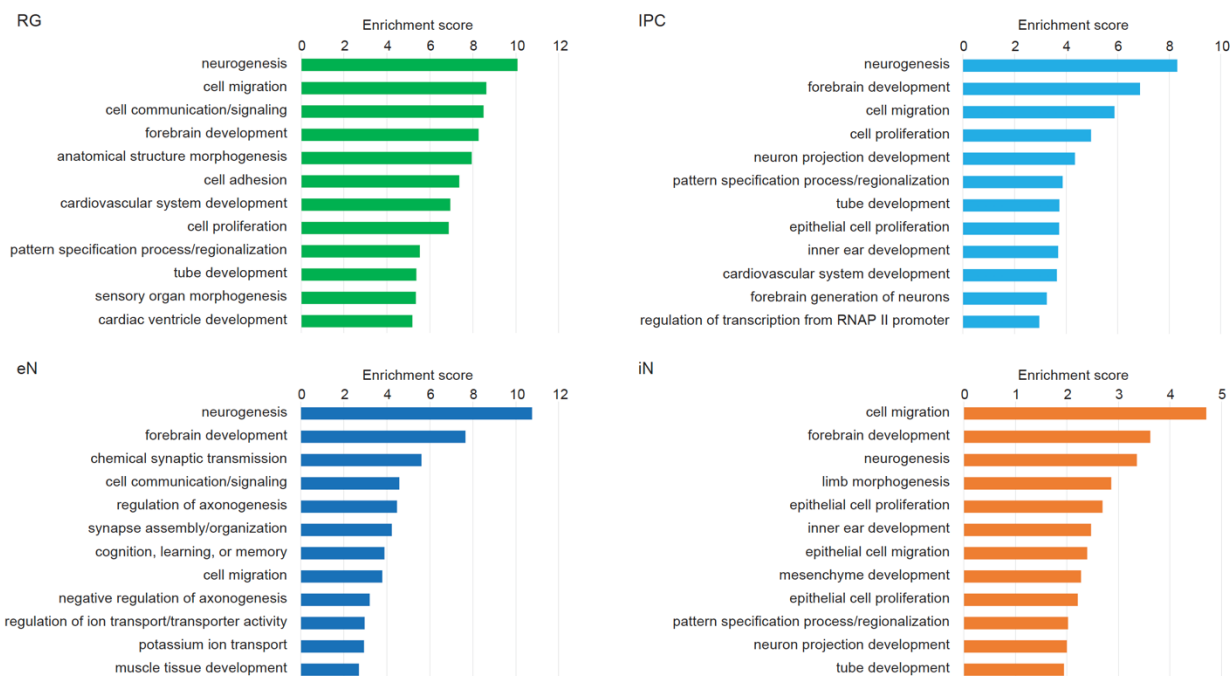
**Figure 2.2. Sorting specific cell types from the developing human cortex.** Within the dorsal cortex, the germinal zone (GZ) is populated by radial glia (RG), which extend fibers towards the cortical plate (CP). These RG divide asymmetrically to produce intermediate progenitor cells (IPCs), which differentiate into excitatory neurons (eNs) that migrate towards the CP. At the same time, interneurons (iNs) can be found in both the GZ and CP. Microdissected GZ and CP samples were dissociated into single cells before being fixed, stained with antibodies for EOMES, SOX2, PAX6, and SATB2, and sorted using FACS.

## 2.2 Characterizing cell type-specific 3D epigenomes

We used H3K4me3 proximity ligation-assisted ChIP-seq (PLAC-seq) (Fang et al., 2016) to identify chromatin interactions at active promoters and assay for transposase-accessible chromatin using sequencing (ATAC-seq) to profile open chromatin peaks for the sorted cell populations (**Figure 2.3c**). After confirming that the samples cluster by cellular identity (**Figures 2.1b-c**), we applied the Model-based Analysis of PLAC-seq (MAPS) pipeline (Juric et al., 2019) to call significant H3K4me3-mediated chromatin interactions at a resolution of 5 kb. We identify 35,552, 26,138, 29,104 and 22,598 interactions in RG, IPCs, eNs, and iNs, respectively, with approximately 85% of the interactions classified as anchor to non-anchor, and the remaining interactions classified as anchor to anchor (**Figure 2.3d**). The median interaction distance was between 170 kb to 230 kb (**Figure 2.3e**), with an average of 4-5 interactions per promoter (**Figure 2.3f**), and the majority of interactions occurred within topologically associated domains (TADs) in the GZ or CP (Won et al., 2016).



**Figure 2.3. Features of 3D epigenomes during human corticogenesis. (a)** Heatmap displaying the expression of key marker genes for each cell type. **(b)** Heatmap showing correlations between gene expression profiles for the sorted cell populations and single-cell RNA sequencing (scRNA-seq) data in the developing human cortex. The sorted cell populations exhibited the highest correlation with their corresponding subtypes while exhibiting reduced correlation with the endothelial, mural, microglial, and choroid plexus lineages. **(c)** WashU Epigenome Browser snapshot displaying a region (chr17: 72,970,000-73,330,000) with interactions linked to *SSTR2* expression in IPCs. **(d)** Bar graph of interaction counts for each cell type, with the proportions of anchor to anchor (red) and anchor to non-anchor (blue) interactions highlighted. **(e)** Cumulative distribution function (CDF) plots of interaction distances for each cell type. **(f)** Histogram displaying the numbers of interactions for interacting promoters across all cell types.

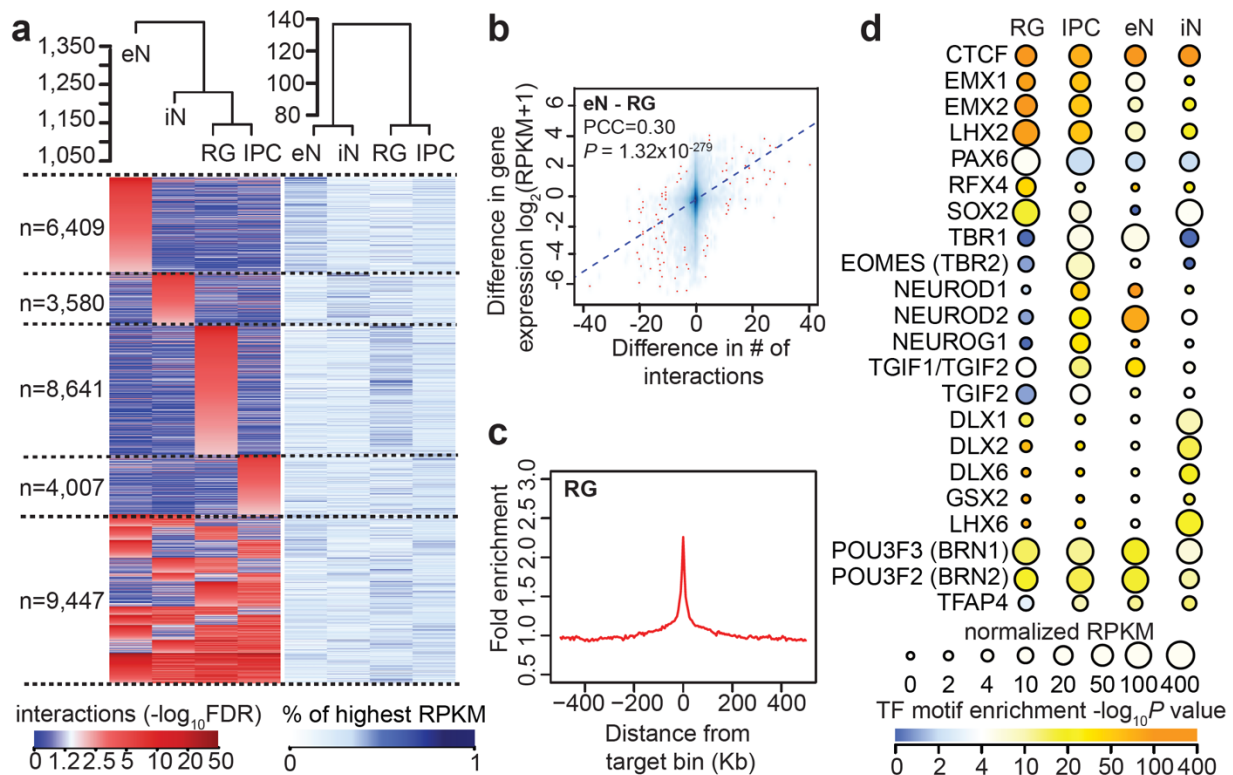


**Figure 2.4. GO enrichment analysis of cell type-specific chromatin interactions.** The top annotation clusters from DAVID are reported along with their group enrichment scores.

## 2.3 Chromatin interactions influence cell type-specific transcription

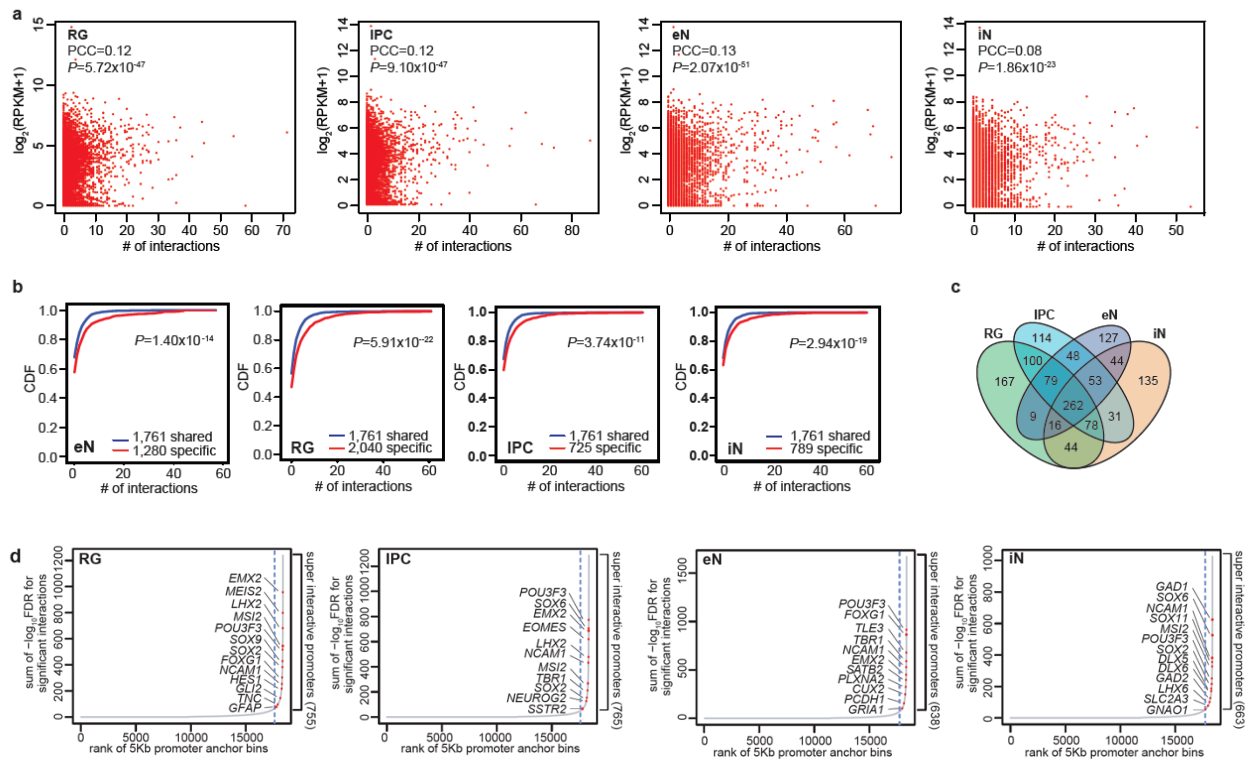
Since H3K4me3 is a mark associated with active promoters, we characterized the extent to which H3K4me3-mediated chromatin interactions influence cell type-specific transcription. First, the sorted cell populations cluster by developmental age based on their interaction strengths across all interacting loci (**Figure 2.5a**). This is consistent with iNs at this age possessing progenitor-like characteristics including high *SOX2* expression. Meanwhile, genes participating in cell type-specific interactions are enriched for biological processes linked to their respective cell types, including cell proliferation for RG and IPCs, neuron projection development for IPCs and eNs, and synaptogenesis for eNs (**Figure 2.4**). Interaction strength and gene expression are positively correlated across all cell types (**Figure 2.5b**), suggesting that chromatin interactions orchestrate transcription in a manner that is distinctly cell type-specific. Next, we leveraged the enrichment of

open chromatin peaks at distal interacting regions (**Figure 2.5c**) and performed transcription factor (TF) motif enrichment analysis for distal interacting regions in each cell type (**Figure 2.5d**). The motifs for PAX6, EOMES, and TBR1 are enriched in RG, IPCs, and eNs, respectively, recapitulating their sequence of expression along this developmental trajectory (Englund et al., 2005). Meanwhile, the motifs for DLX1, DLX2, DLX6, GSX2, and LHX6 are enriched in iNs, in accordance with their roles in iN maturation and function (Lim et al., 2018). Our results link key lineage-specific TFs while linking them to their interacting genes, enabling novel insights into gene regulatory networks during human corticogenesis.



**Figure 2.5. H3K4me3-mediated chromatin interactions influence cell type-specific transcription.** (a) Heatmaps showing interaction strengths (left) and gene expression (right) for anchor to non-anchor interactions grouped according to their cell type specificity. Interaction strengths are based on the  $-\log_{10}$ FDR from the MAPS pipeline. (b) Scatterplot showing the correlation between the difference in the number of interactions for each promoter and the difference in the expression of the corresponding genes for RG and eNs (Pearson product-

moment correlation coefficient, two-tailed,  $P=1.32 \times 10^{-279}$ ,  $n=13,996$  anchor bins with promoters). The trendline from linear regression is shown. **(c)** Fold enrichment of open chromatin peaks over distance-matched background regions in 1 Mb windows around distal interacting regions for RG. **(d)** TF motif enrichment analysis for open chromatin peaks at cell type-specific distal interacting regions in each cell type. Colors represent enrichment scores based on the p-value from HOMER, while sizes represent the gene expression of the corresponding TFs.

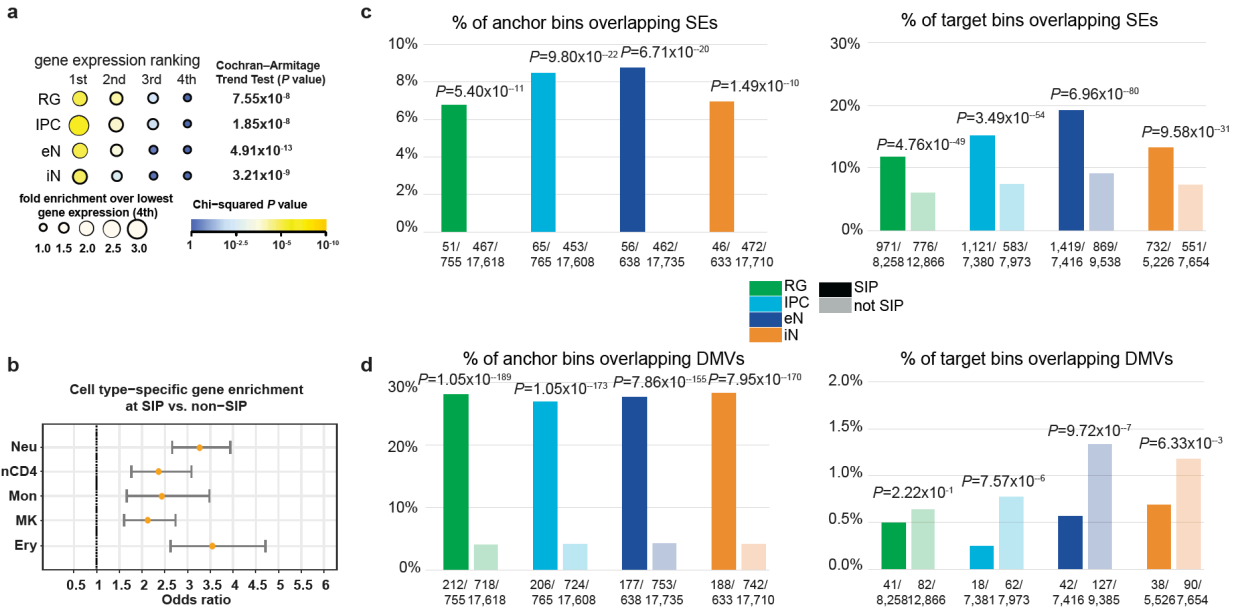


**Figure 2.6. SIPs are enriched for lineage-specific genes.** **(a)** Scatterplots showing the correlation between interaction counts and gene expression at promoters for each cell type (Pearson product-moment correlation coefficient, two-tailed,  $n=13,996$  anchor bins with promoters). **(b)** CDF plots of the numbers of interactions for shared versus cell type-specific genes for each cell type (two-sample t-test, two-tailed). **(c)** Venn diagram displaying cell type-specificity for SIPs in each cell type. **(d)** Anchor bins were ranked according to their cumulative interaction scores in RG, IPCs, eNs, and iNs. Super interactive promoters (SIPs) are located past the point in each curve where the slope is equal to 1.

## 2.4 Super interactive promoters are enriched for lineage-specific genes

The number of chromatin interactions at H3K4me3-mediated anchor bins is only modestly correlated with gene expression (**Figure 2.6a**). One potential explanation is that individual genes are expressed to varying degrees in the contexts of their diverse cellular functions, and a subset of regulatory elements may be better described as fine-tuning rather than independently inducing or silencing the expression of their interacting genes. Multiple regulatory interactions can also exert synergistic or nonlinear effects on transcription. We first demonstrate that cell type-specific genes tend to harbor more chromatin interactions than shared genes across all four cell types (**Figure 2.6b**). Next, by ranking anchor bins according to their cumulative interaction scores, we delineate a subset of promoters with significantly increased levels of chromatin interactivity, termed super interactive promoters (SIPs). We identify 755, 765, 638, and 663 SIPs in RG, IPCs, eNs, and iNs, respectively (**Figures 2.6c-d**). SIPs are enriched for key lineage-specific genes including *GFAP* and *HES1* for RG, *EOMES* for IPCs, *SATB2* for eNs, and *DLX5*, *DLX6*, *GAD1*, *GAD2*, and *LHX6* for iNs. We also observe forebrain-specific SIPs including *FOXP1* in all four cell types, progenitor-specific SIPs including *SOX2* in RG, IPCs, and iNs, and cortical neuron-specific SIPs including *TBR1* in IPCs and eNs. Numerous promoters for lincRNAs including *LINC00461* and *LINC01551* are annotated as SIPs, consistent with their expression in the developing human cortex (Liu et al., 2016). In general, SIPs are enriched in cell types with the highest expression of their linked genes, supporting their putative roles in lineage specification (**Figure 2.7a**). Moreover, super-enhancers and DNA methylation valleys (DMVs) (Luo et al., 2016), two epigenetic features associated with developmental genes and cellular identity, both exhibit enrichment at SIPs (**Figures 2.7c-d**). Finally, SIPs based on promoter capture Hi-C data in neutrophils, naive CD4<sup>+</sup> T cells, monocytes, megakaryocytes, and erythroblasts (Javierre et al., 2016) are analogously enriched for cell type-specific over shared genes (**Figure 2.7b**),

implying that SIPs present a generalized mechanism for maintaining the expression of key genes underlying cellular identity and function.

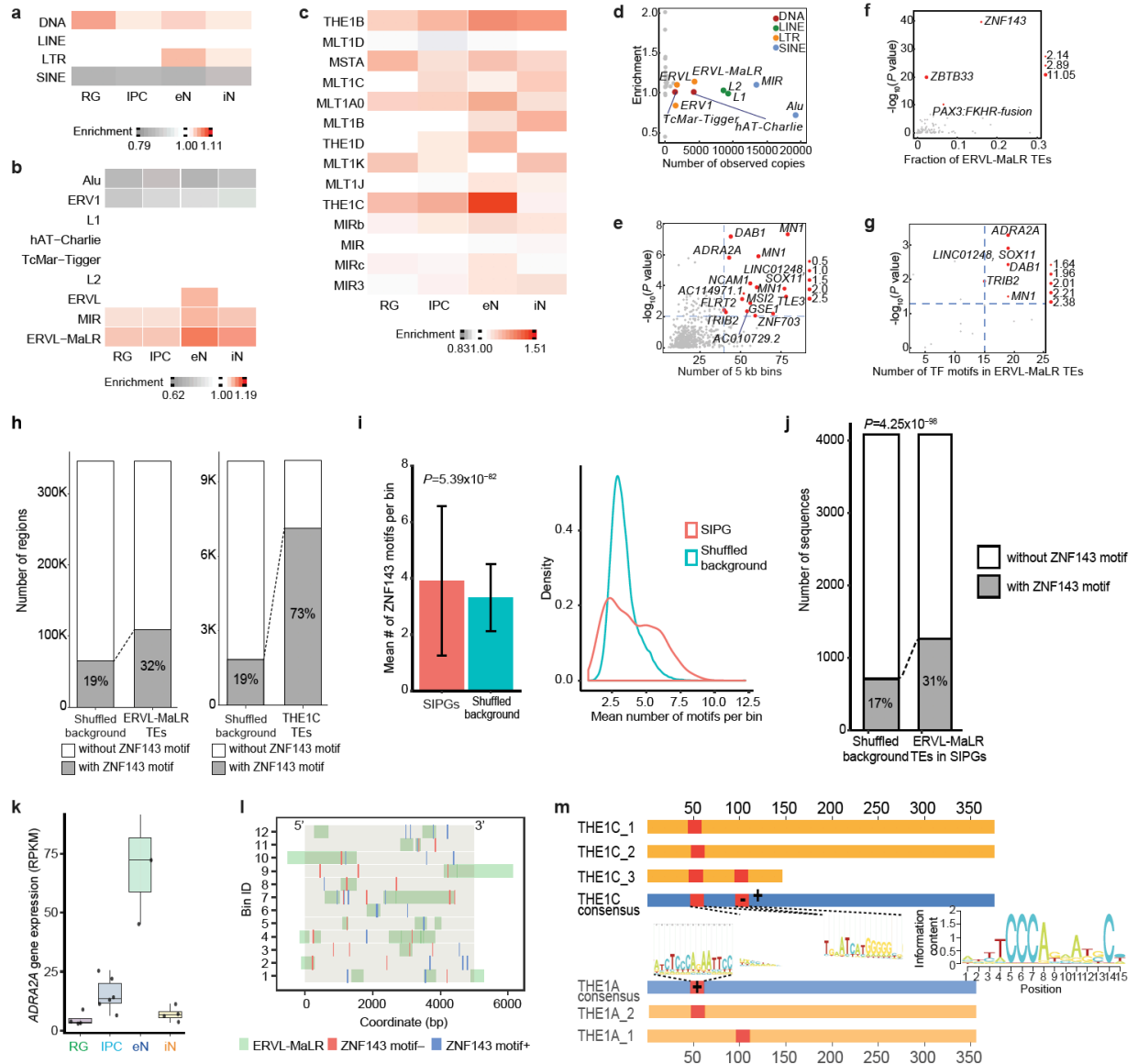


**Figure 2.7. Features of SIPs.** (a) The number of SIPs was divided by the total number of anchor bins (both SIPs and non-SIPs) associated with genes with the 1st, 2nd, 3rd, and 4th highest expression among all four cell types ( $n=13,996$  anchor bins with promoters). Fold enrichment was calculated relative to the group with the lowest expression among all four cell types. (b) Forrest plot showing that SIPs identified in hematopoietic cells are analogously enriched for cell type-specific over shared genes. Odds ratios and 95% confidence intervals are shown. We identified 554, 709, 460, 712, and 401 SIPs in neutrophils, naive CD4<sup>+</sup> T cells, monocytes, megakaryocytes, and erythroblasts, respectively. (c-d) Enrichment of super-enhancers and DMVs at SIPs versus non-SIPs (left) and distal interacting regions for SIPs versus non-SIPs (right) (Fisher's exact test, two-tailed). Super-enhancers were based on data in the fetal brain and adult cortex, while DMVs were based on data in 40 and 60 day cerebral organoids with closely matched gene expression profiles to mid-fetal cortex samples.

## 2.5 Transposable elements in SIP formation

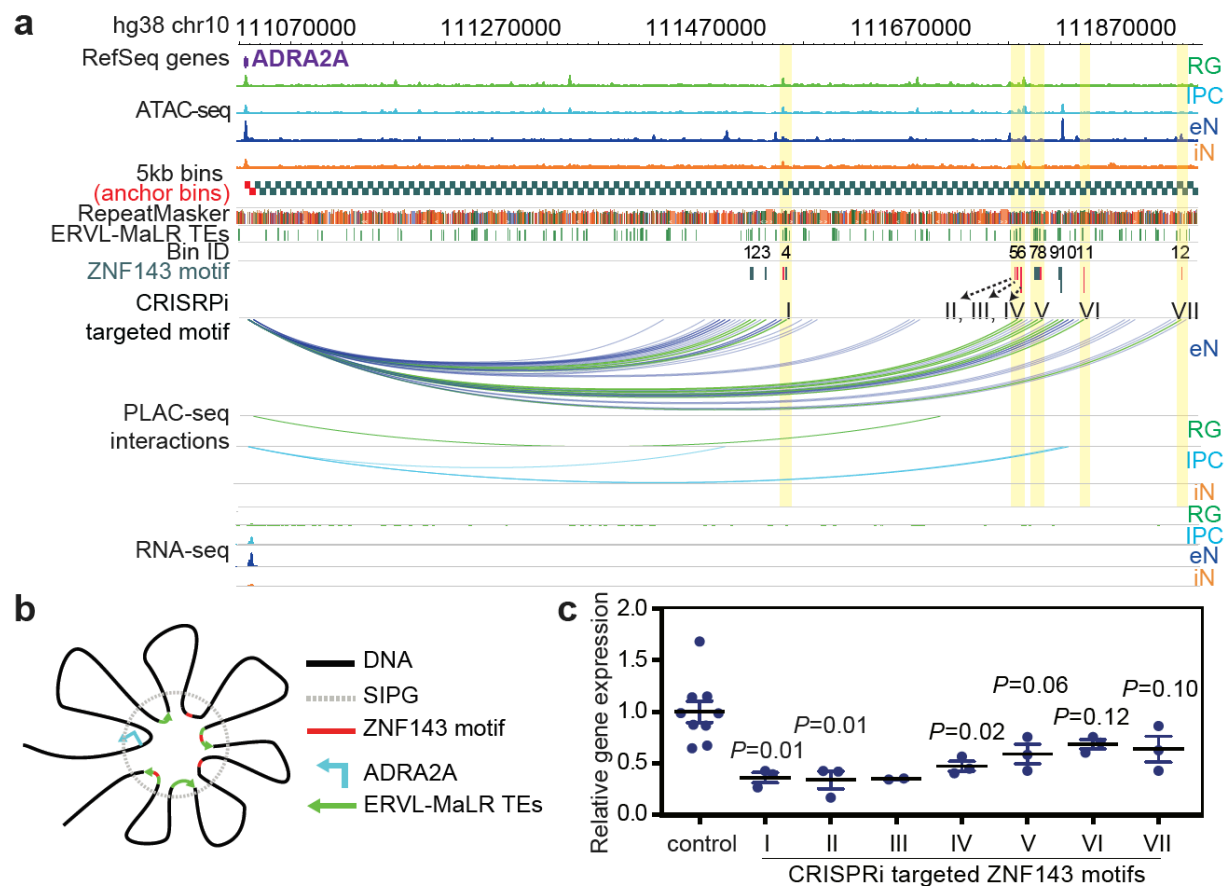
To explore potential mechanisms underlying SIP formation, we evaluated the contributions of transposable elements (TEs), which have been shown to influence 3D chromatin architecture and propagate regulatory elements throughout the genome (Feschotte et al., 2008; Zhang et al., 2019;

Choudhary et al., 2020). We analyzed the enrichment of TEs at the class, family, and subfamily levels in sequences defined by SIPs and their distal interacting regions, termed super interactive promoter groups (SIPGs) (**Figures 2.8a-d**). We first observe that ERVL-MaLR TEs are enriched in SIPGs across all four cell types. Next, we identify 16 SIPGs in eNs that exhibit significant enrichment for ERVL-MaLR TEs and have 40 or more distal interacting regions (hypergeometric test, one-tailed,  $P < 0.01$ ) (**Figure 2.8e**). TF motif enrichment analysis for ERVL-MaLR TEs in these SIPGs reveals the most enriched motif to be that for ZNF143, an architectural protein that has been demonstrated to mediate physical chromatin looping between promoters and distal regulatory elements (Bailey et al., 2015) (**Figure 2.8f**). Moreover, ERVL-MaLR TE subfamilies have been linked to ZNF143 binding in 3T3 and HeLa cells (Ngondo-Mbongo et al., 2013). We find that ZNF143 motifs are broadly enriched in ERVL-MaLR TEs, SIPGs, and ERVL-MaLR TEs in SIPGs (**Figures 2.8h-j**). The *ADRA2A* SIPG in particular exhibits the strongest enrichment of ERVL-MaLR TE-localized ZNF143 motifs (hypergeometric test, one-tailed,  $P=1.59 \times 10^{-6}$ ) (**Figure 2.8g**) and is linked to elevated *ADRA2A* expression in eNs (**Figure 2.8k**). The *ADRA2A* SIPG spans 42 distal interacting regions, 25 of which contain ERVL-MaLR TEs, and 12 of which contain ERVL-MaLR TE-localized ZNF143 motifs (**Figures 2.8l and 2.9a**). Many of these ZNF143 motifs can be mapped back to the consensus sequences of their corresponding ERVL-MaLR TE subfamilies (**Figure 2.8m**). This supports a model in which ZNF143 motifs are coordinately expanded by ERVL-MaLR TE insertion, promoting increased binding site redundancy and strengthened assembly of the *ADRA2A* regulatory unit (**Figure 2.9b**). Using CRISPRi to target ERVL-MaLR TE-localized ZNF143 motifs in the *ADRA2A* SIPG resulted in the significant downregulation of *ADRA2A* expression for 3 of 7 regions in eNs (two-sample t-test, two-tailed,  $P < 0.05$ ) (**Figure 2.9c**), implying that TEs are capable of mediating the formation of higher order chromatin features including SIPs (Sundaram et al., 2018).



**Figure 2.8. Transposable elements in SIP formation.** (a-c) Enrichment of TEs at the class (a), family (b), and subfamily (c) levels in SIPGs for each cell type. Only TE families occupying more than 1% of the genome are shown in (b). Only TE subfamilies from the MIR and ERVL-MaLR TE families occupying more than 0.1% of the genome are shown in (c). (d) Scatterplot showing the enrichment and numbers of observed copies for TE families in SIPGs for eNs. TE families occupying more than 1% of the genome are colored. (e) Scatterplot showing the enrichment and numbers of distal interacting regions for ERVL-MaLR TEs in SIPGs for eNs (n=638 SIPGs). The 16 SIPGs with significant enrichment (hypergeometric test, one-tailed,  $P < 0.01$ ) and 40 or more distal interacting regions are highlighted. (f) Scatterplot showing the enrichment of TF motifs in ERVL-MaLR TEs for the 16 SIPGs highlighted in (e). Enrichment P values are from HOMER. (g) Scatterplot showing the enrichment of ZNF143 motifs in ERVL-MaLR TEs for the 16 SIPGs highlighted in (e) (Poisson distribution, see methods). (h) Both ERVL-MaLR TEs (left, 32% versus 19% of sequences,  $P < 2.2 \times 10^{-16}$ , binomial test, two-tailed) and THE1C TEs (right, 73% versus

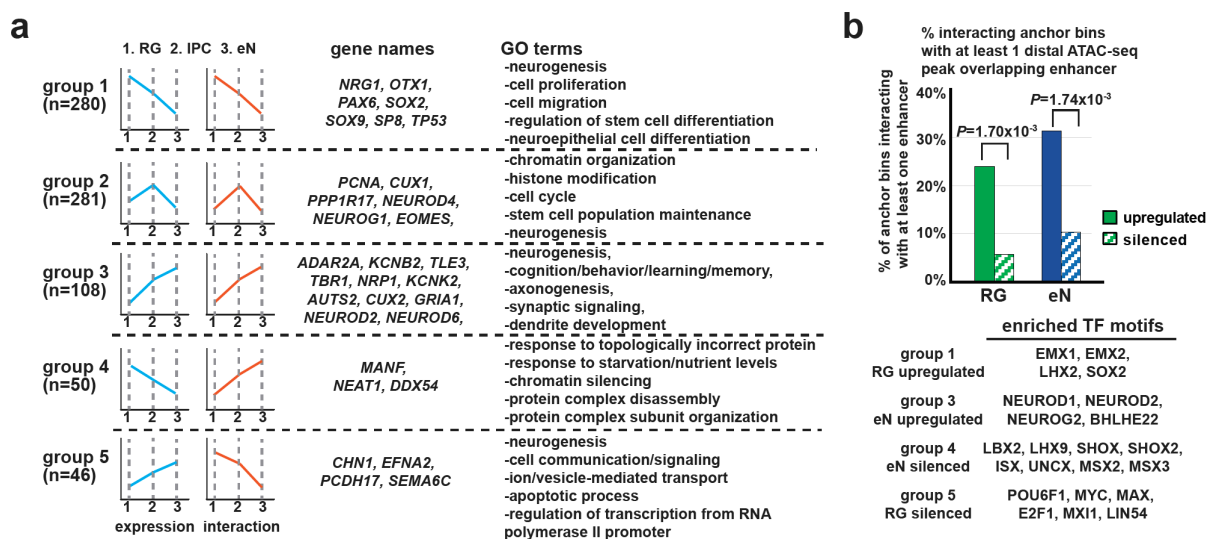
19% of sequences,  $P < 2.2 \times 10^{-16}$ , binomial test, two-tailed) are enriched over background sequences for ZNF143 motifs in eNs. **(i)** ZNF143 motifs are enriched at SIPGs in eNs (left,  $P=5.39 \times 10^{-82}$ , two-sample t-test, two-tailed,  $n=8,894$  distal interacting regions). Means are indicated and error bars represent the SEM. Distributions comparing the number of ZNF143 motifs per bin for actual versus shuffled SIPGs are shown (right,  $P < 2.2 \times 10^{-16}$ , Kolmogorov-Smirnov test, two-tailed,  $n=638$  SIPGs). **(j)** ERVL-MaLR TEs in SIPGs are enriched over background sequences for ZNF143 motifs in eNs (31% versus 17% of sequences,  $P=4.3 \times 10^{-98}$ , binomial test, two-tailed). **(k)** Boxplots showing elevated *ADRA2A* gene expression in eNs. The median, upper and lower quartiles, minimum, and maximum are indicated. **(l)** Illustration of the 12 distal interacting regions containing ERVL-MaLR TE-localized ZNF143 motifs in the *ADRA2A* SIPG. ZNF143 motifs are colored by strand. The bin numbers correspond to Fig. 3j. **(m)** Conservation of ERVL-MaLR TEs in the *ADRA2A* SIPG. Blue bars indicate consensus sequences, yellow bars indicate ERVL-MaLR TEs, and red bars indicate ZNF143 motifs.

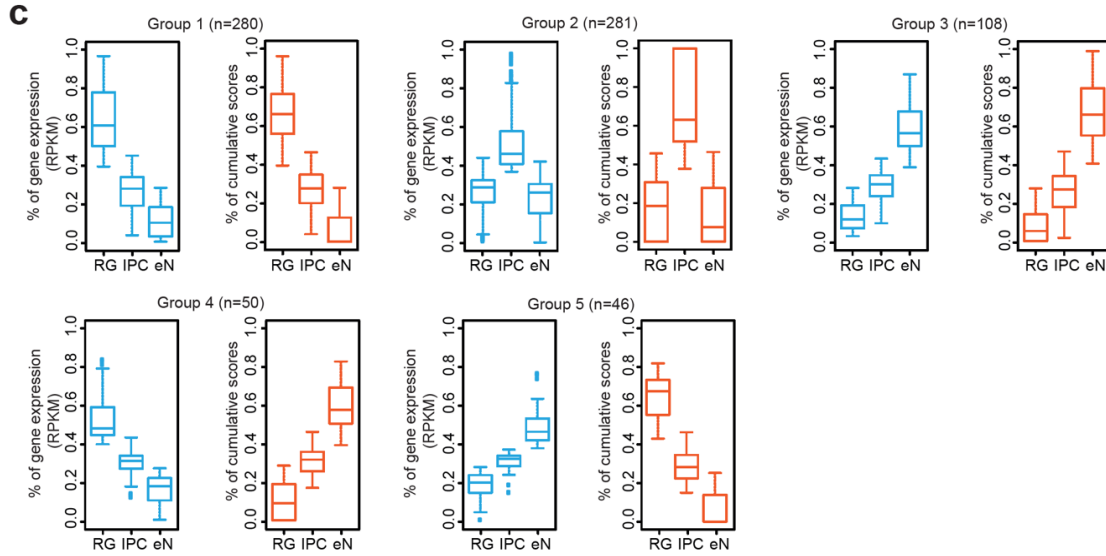


**Figure 2.9. Validating ERVL-MaLR-localized ZNF143 motifs.** **(a)** Interactions between the *ADRA2A* promoter and 12 distal interacting regions containing ERVL-MaLR TE-localized ZNF143 motifs. **(b)** Proposed mechanism for the contribution of TEs to SIP formation. **(c)** *ADRA2A* expression was significantly downregulated for 3 of 7 regions relative to control sgRNAs (two-sample t-test, two-tailed,  $P < 0.05$ ,  $n=3$  for all regions except region III, which has  $n=2$ ). Means are indicated and error bars represent the SEM.

## 2.6 Developmental trajectories from RG to eNs

Since RG, IPCs, and eNs represent a developmental trajectory from dorsal cortical progenitors to mature functional neurons, we grouped genes based on their gene expression and chromatin interactivity along this axis and identified genes linked to cell type-specific processes in RG, IPCs, and eNs (groups 1-3) (**Figures 2.10a** and **2.10c**). We similarly identified genes with anticorrelated gene expression and chromatin interactivity from RG to eNs (groups 4-5), which represent eN-silenced and RG-silenced genes, respectively. eN-silenced genes are enriched for biological processes linked to chromatin remodeling and epigenetic regulation, while RG-silenced genes are enriched for eN-specific signatures. Furthermore, genes in these two groups are depleted for interactions with enhancers annotated using ChromHMM in the germinal matrix (Davis et al., 2018) while exhibiting enrichment for interactions with TFs containing domains associated with transcriptional repression (**Figure 2.10b**). Our results demonstrate that cell type-specific 3D epigenomes are capable of identifying distinct modes of epigenetic regulation during development.



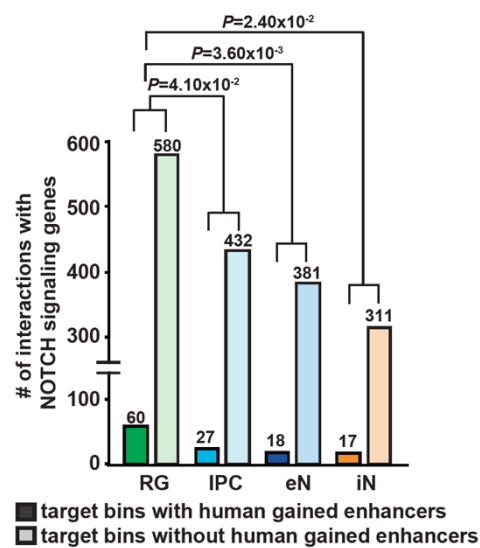


**Figure 2.10. Developmental trajectories from RG to eNs. (a)** Genes categorized based on their gene expression and chromatin interactivity from RG to eNs. Groups 1-5 represent RG-upregulated, IPC-upregulated, eN-upregulated, eN-silenced, and RG-silenced genes, respectively. Representative genes and biological processes are shown for each group. **(b)** Groups 1 (75 of 312 bins) and 3 (40 of 127 bins) are enriched for interactions with enhancers relative to groups 4 (6 of 58 bins) and 5 (3 of 52 bins) (chi-squared test, two-tailed). Only bins with at least one interaction were considered. **(c)** Boxplots showing the distributions of gene expression and cumulative interaction scores for the groups identified in **(a)**. The median, upper and lower quartiles, minimum, and maximum are indicated.

## 2.7 Human-specific aspects of cortical development

Human corticogenesis is dramatically distinct from other mammals, driven largely by the increased diversity and proliferative capacity of cortical progenitors which contribute to the enhanced size and complexity of the human brain (Miller et al., 2019). Notch signaling genes in particular have been implicated in the clonal expansion of RG, which represent the major subtype of cortical progenitors in the developing human cortex (Rani et al., 2016; Suzuki et al., 2018). Here, RG are enriched relative to other cell types for interactions involving Notch signaling genes (Carbon et al., 2009) (**Figure 2.11**). Furthermore, compared to other cell types, interactions in RG target a significantly higher proportion of human-gained enhancers (HGEs) identified through

comparative analyses of human, rhesus macaque, and mouse brains (Reilly et al., 2015). This suggests that epigenetic modifications surrounding Notch signaling genes in RG contribute to significant neurological differences between humans and other species. Additional biological processes exhibiting enrichment for interactions with HGEs include forebrain neuron fate commitment in RG, neuroblast proliferation in IPCs, forebrain neuron development in eNs, and GABAergic interneuron development in iNs.

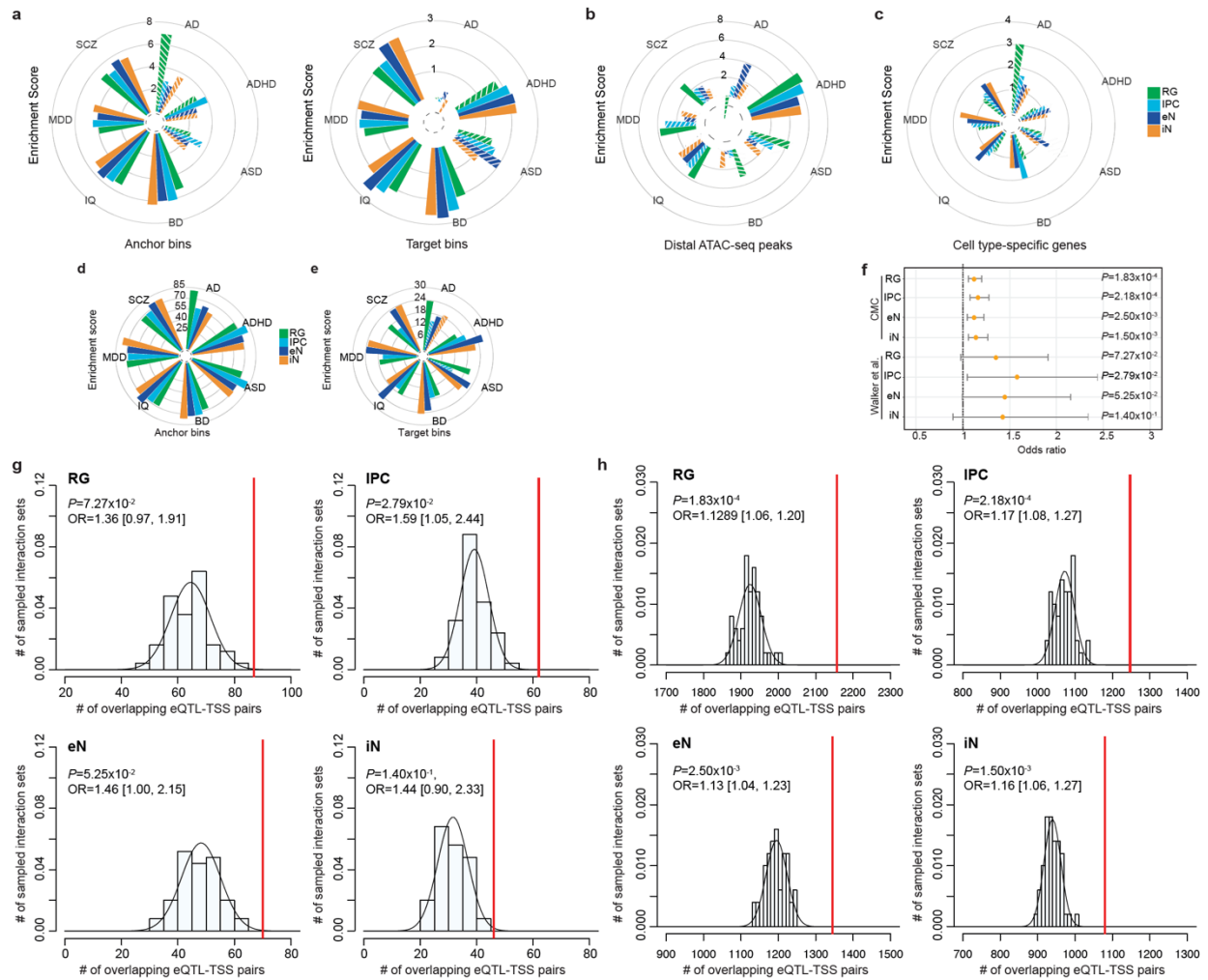


**Figure 2.11. Human-specific aspects of cortical development.** Bar graph of interaction counts from Notch signaling genes to regions with and without HGEs in each cell type (chi-squared test, two-tailed).

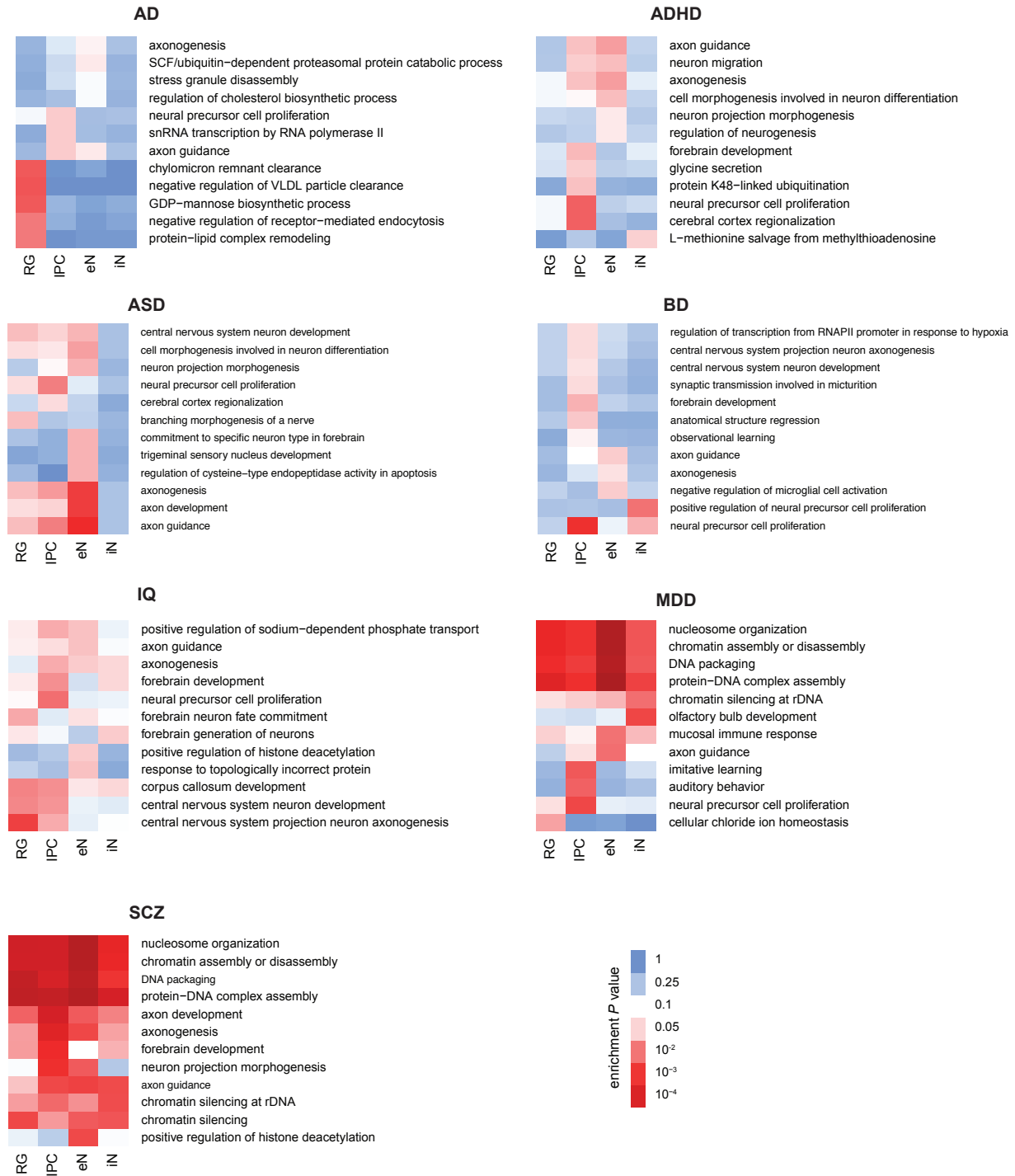
## 2.8 Partitioning SNP heritability for complex disorders and traits

Chromatin interactions present a unique resource for mapping complex disorder- and trait-associated variants to their target genes. This is important as regulatory variants can modulate transcription through the formation or disruption of physical chromatin loops. For instance, both fetal (Walker et al., 2019) and adult (Hoffman et al., 2019) brain eQTLs are enriched at chromatin

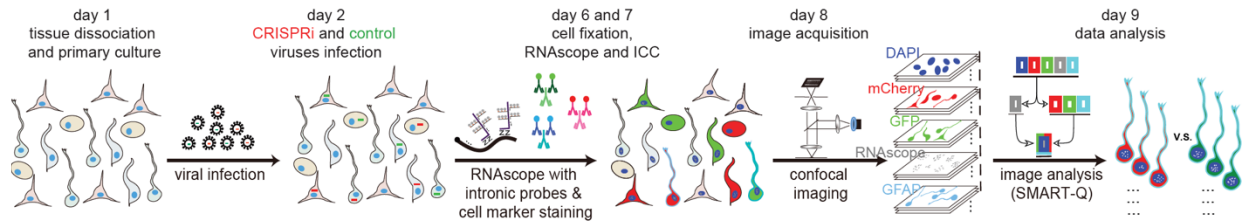
interactions across all four cell types (**Figures 2.12f-h**). Chromatin interactions additionally enable the assessment of cell type-specific patterns of SNP heritability enrichment. Therefore, we leveraged linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015; Finucane et al., 2015) using our 3D epigenomes to partition SNP heritability for seven complex neuropsychiatric disorders and traits: Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD) (Demontis et al., 2019), autism spectrum disorder (ASD) (Grove et al., 2019), bipolar disorder (BD) (Stahl et al., 2019), intelligence quotient (IQ) (Savage et al., 2018), major depressive disorder (MDD) (Howard et al., 2019), and schizophrenia (SCZ) (Pardinas et al., 2018). First, conditioned on a baseline model (Gazal et al., 2017), PLAC-seq anchor and target bins exhibit significant enrichment for all of the disorders and traits we tested, except for AD and ASD (**Figure 2.12a**). Anchor and target bins are also more informative relative to distal open chromatin peaks and cell type-specific genes (**Figures 2.12b-c**). This can be attributed to the utility of chromatin interactions for linking genes to functional regulatory sequences over long genomic distances. Next, we utilized a joint model incorporating all four cell types to investigate cell type-specific patterns of SNP heritability enrichment (**Figures 2.12d-e**). First, target bins exhibit significantly more variability than anchor bins in terms of their enrichment scores, reflecting the increased cell type specificity of distal regulatory elements compared to promoters. Furthermore, eNs and iNs present higher enrichment scores at target bins relative to RG and IPCs, suggesting the increased relevance of neuronal cell types for the disorders and traits in our study. We used H-MAGMA (Sey et al., 2020) to identify enriched biological processes for genes interacting with non-coding variants for each disease and cell type (**Figure 2.13**). Our results recapitulate the roles of lipoprotein metabolism and transport in AD pathophysiology (Andersen et al., 2006). Meanwhile, IPCs and eNs are enriched across all diseases for interactions between SNPs and genes linked to neural precursor cell proliferation, axon guidance, and axonogenesis. Finally, our results for SCZ align with extensive evidence that the disruption of chromatin and epigenetic regulators is a contributor to disease risk (Akbarian et al., 2014; Won et al., 2016).



**Figure 2.12. Partitioning SNP heritability for complex neuropsychiatric disorders and traits.** (a) LDSC enrichment scores for each disease and cell type, conditioned on the baseline model from Gazal et al., 2017 and stratified by PLAC-seq anchor and target bins. Non-significant enrichment scores are shown as striped bars. (b-c) LDSC enrichment scores for each disease and cell type, conditioned on the baseline model from Gazal et al., 2017 and using either distal open chromatin peaks (b) or cell type-specific genes (c). Non-significant enrichment scores are shown as striped bars. (d-e) LDSC enrichment scores for each disease and cell type, stratified by PLAC-seq anchor and target bins. Non-significant enrichment scores are shown as striped bars. (f) Forrest plot showing the enrichment of fetal and adult brain eQTL-TSS pairs in our interactions compared to  $n=50$  sets of distance-matched control interactions (Fisher's exact test, two-tailed). Odds ratios and 95% confidence intervals are shown. The increased significance of adult brain eQTLs can be attributed to the larger sample size of the CommonMind Consortium (CMC) study ( $n=1,332,863$ ), while larger odds ratios were observed for the more closely matched fetal brain eQTLs ( $n=6,446$ ). (g-h) Histograms displaying the numbers of adult and fetal brain eQTL-TSS pairs recapitulated by  $n=50$  sets of distance-matched control interactions in each cell type. The numbers of eQTL-TSS pairs recapitulated by our interactions are indicated by red lines (Fisher's exact test, two-tailed).



**Figure 2.13. GO enrichment analysis of genes interacting with noncoding variants.** GO enrichment analysis was performed for each disease and cell type using H-MAGMA and gProfilerR (Fisher's exact test, two-tailed, BH method). Results are shown for Alzheimer's disease (AD), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BD), intelligence quotient (IQ), major depressive disorder (MDD), and schizophrenia (SCZ).

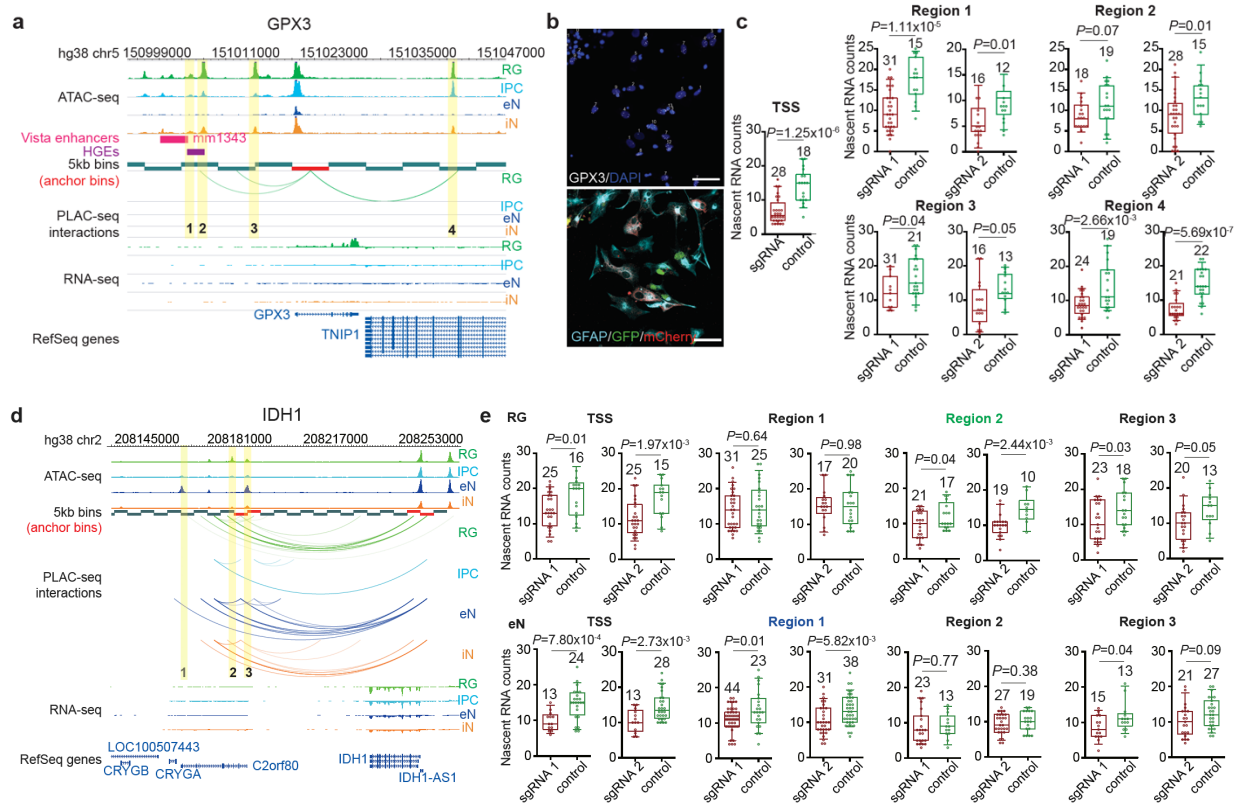


**Figure 2.14. Overview of CRISPRview.** Image analysis was performed using the SMART-Q pipeline.

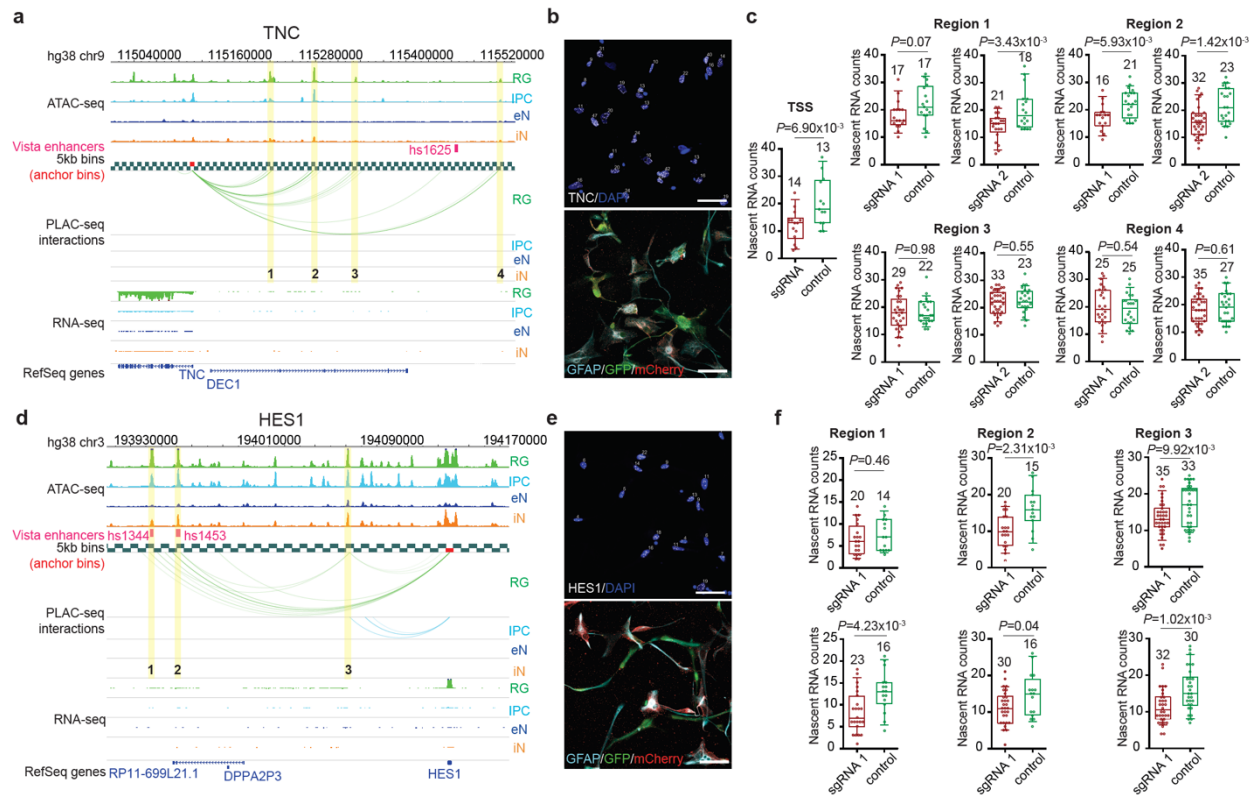
## 2.9 Characterizing distal interacting regions in primary cells

Validating distal regulatory elements in primary cells has proved challenging in the past, with most experiments performed using cell lines or iPSC-derived cells. A major obstacle lies in the robust detection of transcriptional changes resulting from epigenetic perturbations in complex, heterogeneous samples. Therefore, we developed CRISPRview to validate cell type-specific distal regulatory elements in single cells (**Figure 2.14**). Specifically, primary cultures of GZ or CP samples are first infected with lentivirus expressing mCherry, dCas9-KRAB, and sgRNAs targeting open chromatin peaks interacting with a gene of interest, in addition to lentivirus expressing GFP, dCas9-KRAB, and sgRNAs targeting non-human sequences. Next, the cells are fixed and stained using antibodies for mCherry, GFP, cell type-specific markers, DAPI, and intronic RNAscope probes targeting the gene of interest. Finally, we leverage SMART-Q (Yang et al., 2020) to compare the number of nascent RNA transcripts between experimental and control sgRNA-treated cells. We validated four regions interacting with the *GPX3* promoter, all of which exhibited significant downregulation in terms of *GPX3* expression upon silencing (**Figures 2.15a-c**). Meanwhile, silencing three regions interacting with the *IDH1* promoter in RG and eNs resulted in the significant downregulation of *IDH1* expression in the respective cell types (**Figures 2.15d-e**). Finally, we characterized two additional RG-specific loci in *TNC* and *HES1*, both of which are

annotated as SIPs (**Figures 2.16a-f**). The observation of small but significant changes in gene expression supports the hypothesis that multiple interactions frequently work in concert to titrate the expression of key genes underlying cellular identity and function.



**Figure 2.15. Validation of cell type-specific distal regulatory elements using CRISPRview.** (a) Interactions between the *GPX3* promoter and distal interacting regions containing open chromatin peaks that were targeted for silencing are highlighted. Notably, region 1 overlaps both an HGE and Vista enhancer element (mm1343), supporting its function as a putative enhancer. (b) Representative images show staining for intronic RNAscope probes (white), DAPI (blue), GFAP (light blue), GFP (green), and mCherry (red). The scale bar is 50  $\mu$ m. (c) Boxplots show results for experimental (red) and control (green) sgRNA-treated cells for each region (two-sample t-test, two-tailed). The median, upper and lower quartiles, and 10% to 90% range are indicated. Open circles represent single cells. Sample sizes are indicated above each boxplot. (d-e) Validation of distal interacting regions at the *IDH1* locus in RG and eNs. Silencing region 1, which interacts with the *IDH1* promoter only in eNs, results in the significant downregulation of *IDH1* expression in eNs but not in RG. Silencing region 2, which interacts with the *IDH1* promoter only in RG, results in the significant downregulation of *IDH1* expression in RG but not in eNs. Silencing region 3, which interacts with the *IDH1* promoter in both RG and eNs, results in the significant downregulation of *IDH1* expression in both cell types.



**Figure 2.16. Validation of additional RG-specific loci using CRISPRview.** Boxplots show results for experimental (red) and control (green) sgRNA-treated cells for each region (two-sample t-test, two-tailed). The median, upper and lower quartiles, and 10% to 90% range are indicated. Open circles represent single cells. Sample sizes are indicated above each boxplot. **(a-f)** Validation of distal interacting regions at the *TNC* and *HES1* loci in RG. Interactions between the promoters of *TNC* and *HES1* and distal interacting regions containing open chromatin peaks that were targeted for silencing are highlighted. Representative images show staining for intronic RNAscope probes (white), DAPI (blue), GFAP (light blue), GFP (green), and mCherry (red). The scale bar is 50  $\mu$ m.

## 2.10 Discussion

Recent publications leveraging single-cell RNA sequencing have highlighted the heterogeneity of the developing human cortex, with RG, IPCs, eNs, iNs, microglia, endothelial cells, and subplate neurons present in the dorsal cortex alone. Despite significant differences in lineage and maturation state, many of these cell types share intriguing similarities in their transcriptional landscapes. For example, iNs express genes for TFs that are typically associated with RG

proliferation, including *SOX2*, as well as with eN differentiation, including *ASCL1* and *NPAS31*. Therefore, bulk assays cannot be used to reliably distinguish nuanced epigenetic programs driving gene expression in complex tissues. By profiling cell type-specific 3D epigenomic landscapes during human corticogenesis, we not only demonstrate that gene expression is tightly linked to chromatin interactivity, we also identify SIPs that are highly cell type-specific and enriched for key lineage-specific genes. SIPs represent a novel chromatin structural feature that is distinct from A/B compartments (Lieberman-Aiden et al., 2009), TADs (Dixon et al., 2012), and frequently interacting regions (FIREs) (Schmitt et al., 2016). Furthermore, compared to highly interacting regions (HIRs) (Sohby et al., 2019), SIPs are distinct promoter-centric. We uncover a potential mechanism through which specific TE families propagate binding sites for architectural proteins such as ZNF143, facilitating the formation of multi-interaction clusters that may serve to sustain gene expression. While the analysis of TEs is currently constrained by the list of known TF motifs and the resolution of chromatin interactions in this study, future advances should enable the discovery of additional mechanisms through which TEs contribute to 3D chromatin architecture and gene regulation.

Cortical progenitors, eNs, and iNs are highly divergent in terms of their diversity, proliferative capacity, distribution, and functional characteristics between humans and mice. Therefore, important processes occurring during human corticogenesis cannot be fully recapitulated using mouse models. These non-murine features also suggest that enhancer mutations in humans may not adequately phenocopy to mice. Our dataset provides a comprehensive catalog of annotations for human-gained enhancers and complex neuropsychiatric disorder- or trait-associated variants in cell types that are intricately tied to human cortex development, enabling the interpretation and prioritization of regulatory sequences for follow-up studies. Lastly, by developing CRISPRview, we achieve several emergent advantages for validating distal regulatory elements in primary cells. First, we are able to focus our analysis on specific cell types, circumventing averaging effects

associated with bulk measurements in complex samples. Next, we are able to directly compare experimental and control sgRNA-infected cells within the same population. Finally, we achieve enhanced sensitivity and statistical power based on the detection of nascent RNA transcripts in single cells. Future experiments leveraging CRISPRview in live tissue cultures should continue to reveal regulatory relationships in a manner that is truly representative of the complex in vivo environment during human corticogenesis.

## **2.11 Methods**

### **Tissue dissociation**

The tissue dissociation protocol was adapted from Nowakowski et al., 2017. Briefly, samples were first cut into small pieces in artificial cerebrospinal fluid before being added to pre-warmed papain dissociation media (Worthington #LK003150). The samples were incubated in dissociation media for 45 minutes at 37°C. Next, they were triturated, filtered through a 70 µM nylon mesh, and centrifuged for 8 minutes at 300 g. For individual germinal zone (GZ) and cortical plate (CP) cultures, samples were first cut coronally into thin slices. As previously described, cell density drops dramatically past the outer subventricular zone, enabling the clear identification of the outer filamentous zone and subplate. Samples were dissected along this boundary to separate the GZ from the CP prior to dissociation.

### **Sample fixation**

Mid-gestational human cortex samples between GW15 and GW22 were fixed in 2% paraformaldehyde prepared in PBS with gentle agitation for 10 minutes at room temperature. Glycine was added to a final concentration of 200 mM to quench the reactions, and the samples

were centrifuged for 5 minutes at 4°C and 500 g. The samples were washed twice with PBS before being frozen at -80°C for further processing.

### **Permeabilization and staining**

The cell pellet was thawed on ice and resuspended in PBS containing 0.1% Triton-X-100 for 15 minutes. The cells were then washed twice with PBS and resuspended in 5% BSA in PBS for staining. Staining proceeded for at least one hour with FcR Blocking Reagent (Miltenyi Biotech, 1/20 dilution), EOMES PE-Cy7 (Invitrogen, Cat 25-4877-42, Clone WD1928, Lot 1923396, 1/10 dilution), PAX6 PE (BD Biosciences, Cat 561552, Clone O18-1330, Lot 8187686, 1/10 dilution), SOX2 PerCP-Cy5.5 (BD Biosciences, Cat 561506, Clone O38- 678, Lot 8165744, 1/10 dilution), and SATB2 Alexa Fluor 647 (Abcam, Cat ab196536, Clone EPNCIR130A, Lot GR3208103-I and GR228747-2, 1/100 dilution). After staining, the cells were centrifuged for 5 minutes at 500 g, and the pellet was diluted into PBS. When sorting cells for RNA-seq, 1% RNasin Plus RNase Inhibitor (Promega) was added to all buffers, and acetylated BSA was used to prepare 5% BSA in PBS for staining.

### **FACS**

AbC Total Antibody Compensation Beads (Thermo Fisher) were used to generate single color compensation controls prior to sorting. Sorting was conducted on either the FACS Aria II, FACS Aria IIu, or FACS Aria Fusion instruments using a 70 µm nozzle, and cells were collected in 5 ml tubes pre-coated with FBS. A sample of each sorted cell population was reanalyzed on the same machine to assess purity. Cells were collected by centrifuging for 10 minutes at 500 g, and the cell pellet was frozen at -80°C for further processing. When sorting cells for RNA-seq, cells were collected in 5 ml tubes pre-coated with both FBS and RNAlater (Thermo Fisher).

## **Primary cell culture**

Following dissociation, cells were plated onto Matrigel-coated coverslips in 48 well plates or chamber slides at a density of approximately  $0.7 \times 10^6$  cells per well. All cell culture was handled in sterile conditions. The cells were infected with lentivirus the day after plating, and media was changed every two days. Media was composed of 96% DMEM/F-12 with GlutaMAX, 1% N-2, 1% B-27, and 1% penicillin/streptomycin. The cells were grown in 8% oxygen and 5% carbon dioxide and harvested four days post-infection for CRISPRview. For qPCR at the ADRA2A locus, the cells were harvested six days post-infection.

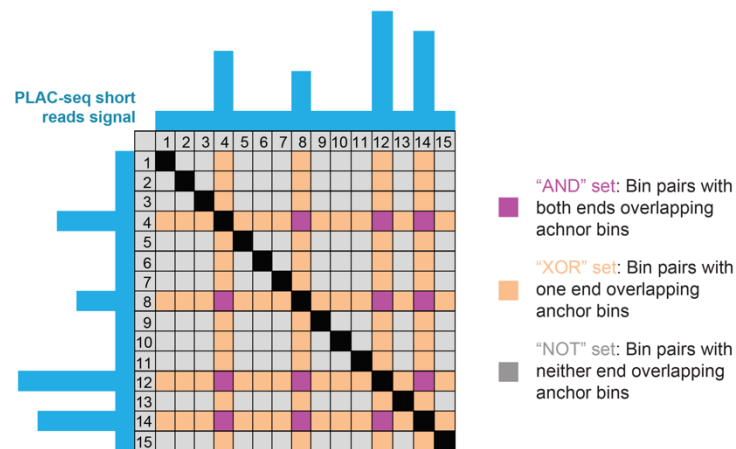
## **PLAC-seq**

PLAC-seq was performed according to Fang et al., 2016<sup>11</sup>. 1 to 5 million cells were used to prepare each library. Digestion was performed using 100 U MboI for 2 hours at 37°C, and chromatin immunoprecipitation was performed using Dynabeads M-280 sheep anti-rabbit IgG (Invitrogen #11203D) superparamagnetic beads bound with 5 µg anti-H3K4me3 antibody (Millipore 04-745). Sequencing adapters were added during PCR amplification. Libraries were sent for paired-end sequencing on the HiSeq X Ten or NovoSeq 6000 instruments (150 bp paired-end reads). fastp was applied to trim reads to 100 bp for all downstream analysis.

## **MAPS**

We used the MAPS pipeline to call significant H3K4me3-mediated chromatin interactions at a resolution of 5 kb based on our PLAC-seq data. First, bwa mem was used to map raw reads to hg38. Unmapped reads and reads with low mapping quality were discarded, and the resulting read pairs were processed as previously reported (Juric et al., 2019). To define PLAC-seq anchor

bins, we took the union of peaks identified by MACS2 using the options “--nolambda --nomodel -extsize 147 --call-summits -B --SPMR” and an FDR cutoff of 0.0001 for all read pairs with interaction distance < 1 kb in each cell type. Next, we classified read pairs as AND, XOR, or NOT interactions based on whether both, one, or neither of the interacting 5 kb bins overlapped anchor bins (**Figure 2.17**). Since we were specifically interested in identifying long-range H3K4me3-mediated chromatin interactions, we retained only read pairs corresponding to intrachromosomal XOR and AND interactions with interaction distances between 10 kb and 1 Mb. We downsampled the number of read pairs separately for each chromosome to ensure that we started with the same number of read pairs for each cell type.



**Figure 2.17. Identifying significant H3K4me3-mediated chromatin interactions.** Illustration of XOR and AND interactions in a representative PLAC-seq contact matrix. Purple cells represent AND interactions where both of the interacting bins are anchor bins. Orange cells represent XOR interactions where only one of the interacting bins is an anchor bin. Grey cells represent NOT interactions where neither of the interacting bins are anchor bins.

To call significant interactions, we employed a Poisson regression-based approach to normalize systematic biases from restriction sites, GC content, sequence repetitiveness, and ChIP enrichment. We fitted models separately for AND and XOR interactions and calculated FDRs for interactions based on the expected and observed contact frequencies between interacting 5 kb

bins. We grouped interactions whose ends were located within 15 kb of each other into clusters and classified all other interactions as singletons. We defined our significant H3K4me3-mediated chromatin interactions as interactions with 12 or more reads, normalized contact frequency (defined as the ratio between the observed and expected contact frequency)  $\geq 2$ , and FDR  $< 0.01$  for clusters and FDR  $< 0.0001$  for singletons. This was based on the reasoning that biologically meaningful interactions are more likely to appear in clusters, while singletons are more likely to represent false positives.

### **ATAC-seq**

ATAC-seq was performed as previously described using the Nextera DNA Library Prep Kit (Illumina #FC-121-1030). Briefly, fixed cells were washed once with ice cold PBS containing 1x protease inhibitor before being resuspended in ice cold nuclei extraction buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 0.1% Igepal CA630, and 1x protease inhibitor) for 5 minutes. 50,000 cells were aliquoted, exchanged into 50  $\mu$ L 1x Buffer TD, and incubated with 2.5  $\mu$ L TDE1 enzyme for 45 minutes at 37°C with shaking. Following transposition, 150  $\mu$ L reverse crosslinking solution (50  $\mu$ L 1 M Tris pH 8.0, 100  $\mu$ L 10% SDS, 2  $\mu$ L 0.5 M EDTA, 10  $\mu$ L 5 M NaCl, 800  $\mu$ L water, and 2.5  $\mu$ L 20 mg/mL Proteinase K) was added to each tube and incubated at 65°C overnight. DNA was column purified, PCR amplified, and size-selected for fragments between 300 and 1000 bp. Libraries were sent for paired-end sequencing on the NovaSeq 6000 instrument (150 bp paired-end reads). Raw reads were trimmed to 50 bp, mapped to hg38, and processed using the ENCODE pipeline ([https://github.com/kundajelab/atac\\_dnase\\_pipelines](https://github.com/kundajelab/atac_dnase_pipelines)) running the default settings. The optimal naive overlap peaks for each cell type were used for all downstream analysis.

## **RNA-seq**

We extracted total RNA from the sorted cell populations using the RNAstorm™ FFPE RNA extraction kit (Cell Data Sciences #CD501) starting with  $5 \times 10^5$  to  $1.5 \times 10^6$  cells. The quality of the extracted RNA was checked by determining the percentage of RNA fragments with size > 200 bp (DV200) from the Agilent 2100 Bioanalyzer. RNA samples with DV200  $\geq 40\%$  were used for library construction. First, samples were depleted of ribosomal RNA using the KAPA RNA HyperPrep Kit with RiboErase (HMR #KK8560). Next, we performed first and second strand synthesis, dA-tailing, and sequencing adapter ligation. cDNA was cleaned up and sequencing adapters were added via PCR amplification. Libraries were sent for paired-end sequencing on the NovaSeq 6000 instrument (150 bp paired-end reads). Raw reads were trimmed using Trim Galore and aligned to hg38 using STAR running the standard ENCODE parameters, and transcript quantification was performed in a strand-specific manner using RSEM with the GENCODE 29 annotation. The edgeR package in R was used to calculate TMM-normalized RPKM values for each gene, and the mean values across all replicates were used for all downstream analysis.

## **GO enrichment analysis**

Protein coding and non-coding RNA genes participating in cell type-specific XOR interactions were used for GO enrichment analysis. Only interactions with open chromatin peaks overlapping promoters (defined as the 1 kb region centered around a gene's TSS) in their anchor bins and distal open chromatin peaks (defined as open chromatin peaks not overlapping promoters) in their target bins were used. A minimum RPKM of 0.5 was used to retain only genes that were expressed, and the resulting genes were input into DAVID 6.8 running functional annotation clustering using the "GOTERM\_BP\_ALL" ontology. Group enrichment scores based on the

geometric mean of EASE scores for terms in each group are reported. To report enriched biological processes for genes interacting with non-coding variants for each disease and cell type, we assigned non-coding SNPs for each disorder and trait to genes based on interactions with the 5 kb bins containing their promoters. Next, we ran H-MAGMA using our annotations to generate ranked lists of gene-level association statistics which were used to perform functional enrichment analysis using the gprofiler2 package in R (Reimand et al., 2007).

### **TF motif enrichment analysis**

We used 200 bp windows centered around open chromatin peaks participating in cell type-specific XOR interactions for TF motif enrichment analysis using HOMER. We used the complete set of vertebrate motifs from the JASPAR database, specifying the “-float” option to adjust the degeneracy threshold, and the entire genome was used as the background. The binomial distribution was used to calculate p-values.

### **Super interactive promoters**

We used an approach similar to calling super-enhancers (Hnisz et al., 2013) to annotate super interactive promoters (SIPs) in each cell type. For each anchor bin, we calculated the cumulative interaction score, defined as the sum of the  $-\log_{10}\text{FDR}$  for interactions overlapping each anchor bin. We used this metric as it accounts for noise and is directly associated with the interaction strength in PLAC-seq data. Next, we prepared plots of ranked cumulative interaction scores for anchor bins in each cell type and defined SIPs to be anchor bins located past the point in each curve where the slope is equal to 1.

## **Cell type-specific versus shared genes**

We classified each gene as cell type-specific or shared according to its Shannon entropy score across all four cell types. Specifically, for each gene, we calculated its relative expression value in each cell type, defined as its RPKM in that cell type divided by the sum of its RPKMs across all four cell types. Next, we calculated the Shannon entropy score for each gene based on its relative expression values across all four cell types. We classified a gene as specific for a cell type if met the following conditions: its Shannon entropy score was  $< 0.01$ , its RPKM was  $> 1$  in that cell type, and its RPKM in that cell type was the highest across all four cell types. All other genes with RPKM  $> 1$  were classified as shared.

## **TE enrichment in SIPGs**

TE enrichment in SIPGs was evaluated as follows. The foreground enrichment was defined as the number of TEs at the class, family, or subfamily levels overlapping SIPGs in each cell type. The background enrichment was defined as the number of TEs overlapping all interacting 5 kb bins (both SIPGs and non-SIPGs). At least 50% of a TE had to overlap a 5 kb bin for it to be considered overlapping. The overall enrichment was defined as the foreground enrichment divided by the background enrichment multiplied by the proportion of interacting 5 kb bins that were assigned to SIPGs.

For the enrichment of SIPGs for ERVL-MaLR TEs, the foreground enrichment for each SIPG was defined as the number of distal interacting regions containing one or more ERVL-MaLR TEs for that SIPG. The background enrichment for each SIPG was defined as the number of randomly shuffled distal interacting regions containing one or more ERVL-MaLR TEs for that SIPG. We computed the background enrichment over 100 permutations. The overall enrichment was

defined as the foreground enrichment divided by the background enrichment. The significance for each SIPG was calculated using the hypergeometric distribution as follows:

$$P = \frac{\binom{m}{q} \binom{n}{k-q}}{\binom{m+n}{k}}$$

where “q” is the number of distal interacting regions containing one or more ERVL-MaLR TEs for that SIPG, “m” is the number of 5 kb bins containing one or more ERVL-MaLR TEs on the same chromosome, “n” is the number of 5 kb bins containing no ERVL-MaLR TEs on the same chromosome, and “k” is the size of the SIPG.

### **ZNF143 motif enrichment**

For the enrichment of SIPGs for ERVL-MaLR TE-localized ZNF143 motifs, the foreground enrichment for each SIPG was defined as the number of ERVL-MaLR TE-localized ZNF143 motifs in its distal interacting regions. FIMO (Grant et al., 2011) was used to detect ZNF143 motifs within ERVL-MaLR TEs. The background enrichment was defined as the total number of ZNF143 motifs in the SIPG. The overall enrichment was defined as the foreground enrichment divided by the background enrichment multiplied by the proportion of the SIPG that is occupied by ERVL-MaLR TEs. The significance for each SIPG was calculated using a Poisson distribution where the number of events (k) is the foreground enrichment and the rate parameter is the background enrichment multiplied by the proportion of the SIPG that is occupied by ERVL-MaLR TEs.

For evaluating the genome-wide enrichment of ZNF143 motifs in ERVL-MaLR and THE1C TEs, we first used FIMO to scan all ERVL-MaLR and THE1C TEs for instances of ZNF143 motifs. As

a background, we scanned 100 sets of chromosome- and length-matched, non-overlapping sequences randomly sampled to avoid gaps and blacklisted regions in the human genome. We used a similar approach to evaluate the enrichment of ZNF143 motifs in ERVL-MaLR TEs in SIPGs. For evaluating the enrichment of ZNF143 motifs in SIPGs, we compared the mean numbers of ZNF143 motifs per 5 kb bin for distal interacting regions across all SIPGs to 100 sets of chromosome- and length-matched, non-overlapping sequences randomly sampled to avoid gaps and blacklisted regions in the human genome. For comparing the distributions of the mean numbers of ZNF143 motifs per 5 kb bin for actual versus shuffled SIPGs, we sampled distal interacting regions for each SIPG 100 times on the same chromosome in a non-overlapping manner.

### **Partitioning SNP heritability for complex disorders and traits.**

We leveraged linkage disequilibrium score regression (LDSC) to partition SNP heritability separately for each complex neuropsychiatric disorder and trait based on joint models incorporating PLAC-seq anchor or target bins across all cell types. We also ran LDSC using a baseline model (Gazal et al., 2017) consisting of coding, UTR, promoter, and intron regions, histone marks, DNase I hypersensitive sites, ChromHMM/Segway predictions, regions that are conserved in mammals, super-enhancers, FANTOM5 enhancers, and LD-related annotations (recombination rate, nucleotide diversity CpG content, etc.) that are not specific to any cell type. This informs us whether our epigenomic annotations for a given cell type are informative for SNP heritability enrichment compared to a comprehensive set of genomic features that has been widely adopted in the field. To compare different epigenomic annotations for each cell type, we used both distal open chromatin peaks and 100 kb windows around the transcription start and end sites of cell type-specific genes based on their Shannon entropy scores and RPKM > 1.

### **Validating ERVL-MaLR-localized ZNF143 motifs**

CRISPRi and qRT-PCR were used to validate ERVL-MaLR TE-localized ZNF143 motifs at distal interacting regions in the ADRA2A SIPG. Of the 12 distal interacting regions containing ERVL-MaLR TE-localized ZNF143 motifs, we were able to design sgRNAs to target ZNF143 motifs overlapping open chromatin peaks for 7 of the regions. ZNF143 motifs were extended by 100 bp in both directions for designing sgRNAs. To maximize CRISPRi efficiency, we designed two sgRNAs for each region and cloned them into the dual expression cassette in the CRISPRi vector as described for CREST-seq (Diao et al., 2017). sgRNA sequences were confirmed by Sanger sequencing and packaged into lentivirus. Primary cell cultures enriched for eNs based on SATB2 staining were infected with lentivirus for 24 hours, and mRNA was extracted on day 7. qRT-PCR was used to quantify *ADRA2A* expression using the following primers: TCGTCATCATCGCCGTGTTC (forward) and AAGCCTTGCCGAAGTACCAG (reverse).

### **Validating distal interacting regions using CRISPRview**

The CRISPRi vector was modified from the Mosaic-seq (Xie et al., 2017) and CROP-seq (Datlinger et al., 2017) vectors. The hU6-sgRNA expression cassette from the CROPseq-Guide-Puro vector (Addgene #86708) was cloned and inserted downstream of the WPRE element in the Lenti-dCas9-KRAB-blast vector (Addgene #89567). The blasticidin resistance gene was replaced with either mCherry or EGFP. sgRNAs targeting open chromatin peaks in distal interacting regions were designed using CHOPCHOP (Labun et al., 2019). Single-stranded DNA was annealed and ligated into the CRISPRi vector at the BsmBI cutting locus. Single clones were picked following transformation, and the sgRNA sequences were confirmed by Sanger sequencing. For lentiviral packaging, the CRISPRi vector, pMD2.G (Addgene #12259), and psPAX (Addgene #12260) were transformed into 293T cells using PolyJet (SigmaGen

Laboratories #SL100688) according to the manufacturer's instructions. Virus-containing media was collected three times over 16 to 20 hours and concentrated using Amicon 10K columns. All lentivirus was immediately stored at -80°C. Primary cell cultures were infected with virus (MOI < 1) 24 hours after plating, and cells were fixed with 4% PFA four days post-infection for FISH and immunostaining.

FISH experiments were performed using the RNAScope Multiplex Fluorescent V2 Assay kit (ACDBio #323100). Probes targeting intronic regions for *GPX3* (ACDBio #572341), *IDH1* (ACDBio #832031), *TNC* (ACDBio #572361), and *HES1* (ACDBio #560881) were custom-designed, synthesized, and labeled with TSA Cyanine 5 (Perkin Elmer #NEL705A001KT, 1:1000 dilution). Fixed cells were pretreated with hydrogen peroxide for 10 minutes and Protease III for 15 minutes, and probes were hybridized and amplified according to the manufacturer's instructions. Slides were washed with PBS before blocking with 5% donkey serum in PBS for 30 minutes at room temperature. Next, slides were incubated with primary antibodies against mCherry (Abcam ab205402, 1/200), GFP (Abcam ab1218, 1/500), and GFAP (Abcam ab7260, 1/400) for RG or SATB2 (Abcam ab92446, 1/300) for eNs overnight at 4°C, followed by incubation with Alexa Fluor 488 donkey anti-mouse IgG (Thermo Fisher Scientific #A21202, 1/800), Alexa-546 nm donkey anti-rabbit IgG (Thermo Fisher Scientific #A10040, 1/500), and Alexa-594 nm goat anti-chicken IgG (Thermo Fisher Scientific #A11042, 1/500) for 1 hour at room temperature. 3D confocal microscopy images were captured using a Leica TCS SP8 with a 40x oil-immersion objective lens (NA=1.30). The z-step size was 0.4 µm. For five color multiplexed imaging, three sequential scans were performed to avoid overlapping spectra. The excitation lasers were 405 nm and 594 nm, 488 nm and 633 nm, and 561 nm. All images were obtained using the same acquisition settings. For FISH analysis, we developed a Python-based pipeline called Single-Molecule Automatic RNA Transcription Quantification (SMART-Q) for quantifying nascent RNA transcripts in single cells. Briefly, the RNAScope channel was first filtered and fitted in three

dimensions using a Gaussian model. Next, segmentation was performed in two dimensions on the DAPI channel to ascertain the location of each nucleus. Finally, segmentation was performed on the remaining channels to identify experimental and control sgRNA-infected RG or eNs for nascent RNA transcript quantification.

## Conclusion

There is a distinct lack of 3D epigenomic annotations in cell types that are relevant to disease and development, especially in the field of brain research. Past studies have relied on heterogeneous tissues comprised of cell types with disparate functions, limiting their abilities to detect and interpret instances of cell type-specific gene regulation. Neurons and glia, for example, represent lineages with divergent functions that coexist in most tissues of the CNS. At the same time, complex diseases often involve multiple dysregulated loci with cell type-specific patterns of activity. This presents unique challenges for deciphering disease etiology, for example in attempting to distinguish causative mechanisms from secondary, reactive phenotypes when multiple cell types are involved. For these reasons, the comprehensive annotation of regulatory relationships in specific, well-characterized cell populations enables the derivation of deeper insights into complex disease biology. Chromatin interactions in particular are ideal for mapping promoters to distal regulatory elements, as they provide direct evidence of regulatory sequences physically contacting loci of interest. To date, several studies have characterized chromatin interactions in fetal brain tissues and cultured neural cells (Won et al., 2016; Rajarajan et al., 2018). However, these studies relied on in situ Hi-C for their interaction calls, which lacks power compared to targeted approaches such as pcHi-C and PLAC-seq.

By performing integrative analysis of promoter-centric, long-range chromatin interactions, open chromatin regions and transcriptomes, we provide comprehensive annotations for promoters and distal promoter-interacting regions (PIRs) in four neural cell types that are highly relevant to complex neuropsychiatric disorders: iPSC-induced excitatory neurons, iPSC-derived hippocampal dentate gyrus (DG)-like neurons, iPSC-induced lower motor neurons, and primary astrocytes. We identify putative gene targets for both in vivo-validated enhancer elements from the VISTA Enhancer Browser as well as disease-associated variants, enabling the functional

validation of PIRs driving diverse processes in cellular identity and disease. Furthermore, we provide epigenomic characterizations for radial glia, intermediate progenitor cells, excitatory neurons, and interneurons isolated from mid-gestational human cortex samples. We show that chromatin interactions underlie multiple aspects of gene regulation, with transcription factor motifs, families of transposable elements, and disease-associated variants enriched at distal interacting regions in a cell type-specific manner. Notably, promoters with significantly increased levels of chromatin interactivity, termed super interactive promoters, are enriched for lineage-specific genes, evidencing that interactions at these loci contribute to the fine-tuning of cell type-specific transcription. Finally, we develop CRISPRview, a novel technique combining immunostaining, CRISPRi, RNAscope, and image analysis for validating cell type-specific cis-regulatory elements in heterogeneous populations of primary cells. Our study presents the first characterization of cell type-specific 3D epigenomes in the aforementioned human neural cell types and the developing human cortex, revealing critical insights into genome organization, gene regulation, and lineage specification during human brain development.

## References

- Abdulla, S. et al. Hippocampal degeneration in patients with amyotrophic lateral sclerosis. *Neurobiol. Aging* 35, 2639-2645 (2014).
- Akbadian, S. Epigenetic mechanisms in schizophrenia. *Dialogues Clin Neurosci* 16, 405-417 (2014).
- Anderson, S., Mione, M., Yun, K. & Rubenstein, J. L. Differential origins of neocortical projection and local circuit neurons: role of Dlx genes in neocortical interneuronogenesis. *Cereb Cortex* 9, 646-654 (1999).
- Andersen, O. M. & Willnow, T. E. Lipoprotein receptors in Alzheimer's disease. *Trends Neurosci* 29, 687-694 (2006).
- Bailey, S. D. et al. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* 2, 6186 (2015).
- Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24, i153-159 (2008).
- Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58 (3), 268-276 (2012).
- Booth, H. D. E., Hirst, W. D. & Wade-Martins, R. The role of astrocyte dysfunction in Parkinson's disease pathogenesis. *Trends Neurosci* 40, 358-370 (2017).
- 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. *Nature Methods* 10 (12), 1213-8 (2013).
- Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291-295 (2015).

- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47, D1005-D1012 (2019).
- Burghes, A. H. & Beattie, C. E. Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick? *Nat Rev Neurosci* 10, 597-609 (2009).
- Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* 17, 127 (2016).
- Carbon, S. et al. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288-289 (2009).
- Chen, J. A. et al. The Emerging Picture of Autism Spectrum Disorder: Genetics and Pathology. *Annu Rev Pathol* 10, 111-44 (2015).
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884-i890 (2018).
- Choudhary, M. N. et al. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol* 21, 16 (2020).
- Chuang, H. C., Huang, T. N. & Hsueh, Y. P. Neuronal excitation upregulates Tbr1, a high-confidence risk gene of autism, mediating Grin2b expression in the adult brain. *Front Cell Neurosci* 8, 280 (2014).
- Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 373, 895-907 (2015).
- Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012).
- Consortium, G. T. et al. Genetic effects on gene expression across human tissues. *Nature* 550, 204-213 (2017).
- Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 14, 297-301 (2017).

- Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 46, D794-D801 (2018).
- de la Torre-Ubieta, L. et al. The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* 172, 289-304 e218 (2018).
- Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10, 5-6 (2013).
- Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* 51, 63-75 (2019).
- Diao, Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* 14, 629-635 (2017).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376-380 (2012).
- Eisenstein, S. A. et al. Prediction of striatal D2 receptor binding by DRD2/ANKK1 TaqIA allele status. *Synapse* 70, 418-431 (2016).
- Englund, C. et al. Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. *J Neurosci* 25, 247-251 (2005).
- Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539, 452-455 (2016).
- Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 9, 215-216 (2012).
- Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* 26, 1345-1348 (2016).
- Fernandopulle, M. S. et al. Transcription Factor-Mediated Differentiation of Human iPSCs into Neurons. *Curr Protoc Cell Biol* 79, e51 (2018).

- Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9, 397-405 (2008).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 47, 1228-1235 (2015).
- Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* 49, 1421-1427 (2017).
- Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017-1018 (2011).
- Grimmler, M. et al. Unrip, a factor implicated in cap-independent translation, associates with the cytosolic SMN complex and influences its intracellular localization. *Hum Mol Genet* 14, 3099-3111 (2005).
- Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* 51, 431-444 (2019).
- Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162, 900-910 (2015).
- Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 43, 630-638 (2011).
- Hansen, D. V., Lui, J. H., Parker, P. R. & Kriegstein, A. R. Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature* 464, 554-561 (2010).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38, 576-589 (2010).
- Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* 155, 934-947 (2013).
- Hoffman, G. E. et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data* 6, 180 (2019).

- Hou, C., Dale, R. & Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci USA* 107, 3651-3656 (2010).
- Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* 22, 343-352 (2019).
- Huang, C. W. et al. Conditional knockout of breast carcinoma amplified sequence 2 (BCAS2) in mouse forebrain causes dendritic malformation via beta-catenin. *Sci Rep* 6, 34927 (2016).
- Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO J* 12, 1059-1065 (1993).
- Javierre, B. M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* 167, 1369-1384 e1319 (2016).
- Jung, I., Schmitt, A., Diao, Y. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* 51, 1442-1449 (2019).
- Juric, I. et al. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput Biol* 15, e1006982 (2019).
- Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 44, W90-97 (2016).
- Labun, K. et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* 47, W171-W174 (2019).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359 (2012).
- Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213-2233 (2003).

- Li, Y., Hu, M. & Shen, Y. Gene regulation in the 3D genome. *Hum Mol Genet* 27, R228-R233 (2018).
- Liddell, J. R. Are astrocytes the predominant cell type for activation of Nrf2 in aging and neurodegeneration? *Antioxid* 6, E65 (2017).
- Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289-293 (2009).
- Lim, L., Mi, D., Llorca, A. & Marin, O. Development and Functional Diversification of Cortical Interneurons. *Neuron* 100, 294-313 (2018).
- Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* 17, 67 (2016).
- Luo, C. et al. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep* 17, 3369-3384 (2016).
- Lydall, G. J. et al. Confirmation of prior evidence of genetic susceptibility to alcoholism in a genome-wide association study of comorbid alcoholism and bipolar disorder. *Psychiatr Genet* 21, 294-306 (2011).
- Mertens, J. et al. Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder. *Nature* 527, 95-99 (2015).
- Miller, D. J., Bhaduri, A., Sestan, N. & Kriegstein, A. Shared and derived features of cellular diversity in the human cerebral cortex. *Curr Opin Neurobiol* 56, 117-124 (2019).
- Miyaoka, Y. et al. Isolation of single-base genome-edited human iPS cells without antibiotic selection. *Nat Methods* 11, 291-293 (2014).
- Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* 49, 1602-1612 (2017).
- Nadif Kasri, N., Nakano-Kobayashi, A. & Van Aelst, L. Rapid synthesis of the X-linked mental retardation protein OPHN1 mediates mGluR-dependent LTD through interaction with the endocytic machinery. *Neuron* 72, 300-315 (2011).

- Ngondo-Mbongo, R. P., Myslinski, E., Aster, J. C. & Carbon, P. Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11. *Nucleic Acids Res* 41, 4000-4014 (2013).
- Nowakowski, T. J. et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 358, 1318-1323 (2017).
- Pardinas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 50, 381-389 (2018).
- Pontious, A., Kowalczyk, T., Englund, C. & Hevner, R. F. Role of intermediate progenitor cells in cerebral cortex development. *Dev Neurosci* 30, 24-32 (2008).
- Rajarajan, P. et al. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* 362, eaat4311 (2018).
- Rani, N. et al. A Primate lncRNA Mediates Notch Signaling during Neuronal Development by Sequestering miRNA. *Neuron* 90, 1174-1188 (2016).
- Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665-1680 (2014).
- Rees, E., O'Donovan, M. C. & Owen, M. J. Genetics of Schizophrenia. *Current Opinion in Behavioral Sciences, Behavioral Genetics* 2, 8-14 (2015).
- Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 35, W193-200 (2007).
- Ren, G. et al. CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Mol Cell* 67, 1049-1058 e1046 (2017).
- Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317-330 (2015).
- Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* 50, 912-919 (2018).

- Schmitt, A. D. et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* 17, 2042-2059 (2016).
- Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* 20, 437-455 (2019).
- Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31, 1111-1118 (2013).
- Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 16, 259 (2015).
- Sey, N. Y. A. et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci* 23, 583-593 (2020).
- Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* 488, 116-120 (2012).
- Shen, Y. et al. Postnatal activation of TLR4 in astrocytes promotes excitatory synaptogenesis in hippocampal neurons. *J Cell Biol* 215, 719-734 (2016).
- Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6, e21856 (2017).
- Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371-375 (2014).
- Sobhy, H., Kumar, R., Lewerentz, J., Lizana, L. & Stenberg, P. Highly interacting regions of the human genome are enriched with enhancers and bound by DNA repair proteins. *Sci Rep* 9, 4577 (2019).
- Song, M., Yang, X., Ren, X. et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat Genet* 51, 1252-1262 (2019).
- Stahl, E. A. et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet* 51, 793-803 (2019).

- Sudhof, T. C. Calcium control of neurotransmitter release. *Cold Spring Harb Perspect Biol* 4, a011353 (2012).
- Sundaram, V. & Wang, T. Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells: Re-Visiting the "Gene-Battery" Model. *Bioessays* 40 (2018).
- Suzuki, I. K. et al. Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* 173, 1370-1384 e1316 (2018).
- Thomsen, E. R. et al. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat Methods* 13, 87-93 (2016).
- Tiwari, M., Lopez-Cruzan, M., Morgan, W. W. & Herman, B. Loss of caspase-2-dependent apoptosis induces autophagy after mitochondrial oxidative stress in primary cultures of young adult cortical neurons. *J Biol Chem* 286, 8493-8506 (2011).
- Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 35, D88-92 (2007).
- Walker, R. L. et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* 179, 750-771 e722 (2019).
- Wang, C. et al. Scalable production of iPSC-derived human neurons to identify Tau-lowering compounds by high-content screening. *Stem Cell Rep* 9, 1221-1233 (2017).
- Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40, D930-934 (2012).
- Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* 4, 1310 (2015).
- Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 538, 523-527 (2016).
- Xie, Y. H. et al. Cloning and characterization of human IC53-2, a novel CDK5 activator binding protein. *Cell Res* 13, 83-91 (2003).

- Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell* 66, 285-299 e285 (2017).
- Yang, T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* 27, 1939-1949 (2017).
- Yang, X. et al. SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription. *PLoS One* 15, e0228760 (2020).
- Yin, X., Warner, D. R., Roberts, E. A., Pisano, M. M. & Greene, R. M. Novel interaction between nuclear co-activator CBP and the CDK5 activator binding protein - C53. *Int J Mol Med* 16, 251-256 (2005).
- Yu, D. X. et al. Modeling hippocampal neurogenesis using human pluripotent stem cells. *Stem Cell Rep* 2, 295-310 (2014).
- Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 504, 306-310 (2013).
- Zhang, Y. et al. Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* 78, 785-798 (2013).
- Zhang, J. P. et al. Association of a schizophrenia risk variant at the DRD2 locus with antipsychotic treatment response in first-episode psychosis. *Schizophr Bull* 41, 1248-1255 (2015).
- Zhang, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* 89, 37-53 (2016).
- Zhang, Y. et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* 51, 1380-1388 (2019).
- Zhao, X. et al. Caspase-2 cleavage of tau reversibly impairs memory. *Nat Med* 22, 1268-1276 (2016).
- Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* 20, 535-550 (2019).

Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555, 524-528 (2018).

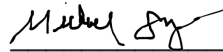
Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 48, 481-487 (2016).

## 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:



571EFCC0B5504A8...

Author Signature

8/24/2020

Date