Supplementary Materials for

Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants

Siwei Zhang[1,†], Hanwen Zhang[1,†], Yifan Zhou[2,†], Min Qiao[2,3,†], Siming Zhao[2,†], Alena Kozlova[1], Jianxin Shi[4], Alan R. Sanders[1,5], Gao Wang[2], Kaixuan Luo[2], Subhajit Sengupta[1], Siobhan West[1], Sheng Qian[2], Michael Streit[1], Dimitrios Avramopoulos[6], Chad A. Cowan[7], Mengjie Chen[8], Zhiping P. Pang[9], Pablo V. Gejman[1,5], Xin He[2,10,*], and Jubao Duan[1,5,*]

† These authors contributed equally to this work.
* Correspondence: xinhe@uchicago.edu (X. H.); jduan@uchicago.edu (J. D.)

This PDF file includes:

Materials and Methods
Figs. S1 to S26
Table Legends S1 to S28

Other Supplementary Materials for this manuscript include the following:

Data Tables S1 to S28 (as a separate file)
Captions for Tables S1 to S28 (included in the Excel file)

Table of Contents

Materials and Methods

1.1. iPSC lines and characterization

The 20 iPSC lines used in open chromatin mapping (ATAC-seq), RNA-seq, and scRNA-seq were reprogrammed from the cryopreserved lymphocytes (CPLs) at Rutgers University Cell and DNA Repository (RUCDR)-NIMH Stem Cell Center. Briefly, we selected from the Molecular Genetics of Schizophrenia (MGS) cohort (with >5,000 schizophrenia [SZ] cases and controls) (*9*) a set of 20 subjects who were enriched for heterozygous SZ GWAS index SNPs at about 70 SZ loci out of 108 SZ risk loci at that time (*10*) (*i.e.*, about 70 GWAS index SNPs are heterozygous); these subjects yielded >80% power to detect ASoC with a high (more frequent) allelic read fraction of 0.65 (*i.e.*, allelic ratio of 1.86) with minimum ATAC-seq read depth of 20 at the targeted 70 loci (Fig. S2) (*10*). These 20 subjects were well balanced in sex and SZ case/control status (Table S1). We found no effect of these traits (sex, age, or disease status) on allele-specific open chromatin (see below the section, "Regression analysis of the association of allelic imbalance with known sample characteristics"). We excluded subjects with large SZ-risk CNVs (*27*) to avoid any confounding effects from such CNVs. The iPSCs were derived from CPLs by using the genome-integration-free Sendai virus method (Cytotune Sendai Virus 2.0; Invitrogen). The iPSC lines were confirmed by immunofluorescence (IF) staining of pluripotency markers (OCT4, SSEA4, NANOG, and TRA-1-60). The pluripotency of the iPSCs was further confirmed by Pluritest with RNA-seq data (*28*). Confirmation of the absence of chromosomal abnormality was performed by eSNP-Karyotyping (Fig. S4; see below) (*29*). The NorthShore University HealthSystem Institutional Review Board (IRB) approved the study.

1.2. Differentiation of iPSCs into neural progenitor cells (NPCs)

NPCs were prepared using PSC Neural Induction Medium (Thermofisher) following the vendor's protocol (except for lines CD0000017 and CD0000020; see below). Briefly, hiPSCs were replated as clumps in mTeSR media (STEMCELL) supplemented with 5 μM ROCK inhibitor (R&D Systems) on Day 0. From Day 1 to Day 9, cells were cultured in PSC Neural Induction Medium with daily media change and routine clean-up of differentiated non-neuronal cells. PSC Neural Induction Medium was prepared as 490 ml Neurobasal Medium (Thermofisher) and 10 ml Neural Induction Supplement (Thermofisher). On Day 10, passage 0 (P0) NPCs were ready for passaging. From P0 to P3, NPCs were passaged using Rosette Selection Reagent (STEMCELL)

and expanded in Neural Expansion Medium (NEM). NEM was prepared as 245 ml Neurobasal Medium, 245 ml Advanced DMEM/F-12 media (Thermofisher), and 10 ml Neural Induction Supplement. Media were changed every other day after switching to NEM. From P4, NPCs were passaged as single cells using Accutase (Thermofisher). 5 µM ROCK inhibitor was supplemented into media on the day of passaging and withdrawn 24 hrs post replating. NPCs were maintained for no more than ten passages.

NPCs for line CD0000017 and CD0000020 were prepared using the embryoid body (EB) method. Briefly, iPSCs were detached as colonies using collagenase IV on Day 0. Colonies were carefully transferred into a low-attachment 6-well plate (Corning) and cultured in EB Medium from Day 0 to Day 4 with daily media change. EB medium was prepared as 390 ml DMEM/F12 Media (Thermofisher), 100 ml Knockout Serum Replacer (Thermofisher), 1x Glutamax (Thmofisher), 1x NEAA, 1x 2-mercaptoethanol, 1 ml Primocin (InvivoGen), 2 µM Dorsomorphin (R&D), and 2 µM A-83 (R&D). From Day 5 to Day 7, the culture media was switched to hNPC media supplemented with 2 µM Dorsomorphin and 2 µM A-83 with daily media change. hNPC media was prepared as 490 ml DMEM/F12, 1x NEAA, 1x N2 supplement, 1ml Primocin, 2 µg/ml Heparin (R&D), and 2 µM cyclopamine (R&D). On Day 7, EBs were transferred into Geltrex-coated 6-well plates to form rosettes. From Day 8 to Day 16, cells were expanded in hNPC media without Dorsomorphin or A-83 and media were changed every other day. On Day 17, rosettes were isolated using the Rosette Selection Reagent and transferred onto Geltrex-coated 6-well plates in StemPro NSC SFM. StemPro NSC SFM was prepared as 484 ml Knockout DEM/F12 (Thermofisher), 10 ml StemPro Neural Supplement, 1x GlutaMAX-I CTS, 1 ml Primocin, 10 µg recombinant Human FGF-basic, and 10 µg Recombinant Human EGF. From Day 17 to Day 24, rosettes were further expanded in StemPro NSC SFM and media was changed every two days. On Day 25, P0 NPCs were single-cell passaged using Accutase in StemPro NSC SFM. 5 µM ROCK inhibitor was supplemented into media on the day of passaging and withdrawn 24 hrs post replating. NPCs were maintained for no more than 10 passages.

1.3. Differentiation into glutamatergic neurons

We generated glutamatergic neurons (iN-Glut) from iPSC-differentiated NPCs using a method adapted from Zhang *et al.* and Wen *et al.* (*30, 31*). Briefly, NPCs were seeded as single cells into Geltrex-coated plates at 40,000 cells/cm^2 on Day 0. From Day 0 to Day 14, NPCs were

cultured in NeuralbasalPlus media and the media was changed every two days. NeuralbasalPlus media was prepared as 490ml NeuralBasal, 1x B27Plus Supplement (Thermofisher), 1x Glutamax, 10 ng/ml BDNF (PeproTech), 10ng/ml GDNF (PeproTech), and 10ng/ml NT-3 (PeproTech). On Day 15, pre-mature neuron cells (iN-Glut) were harvested using Accutase for ATAC-seq or RNA isolation.

1.4. Differentiation into GABAergic neurons

We generated GABAergic neurons (iN-GABA) from NPCs using a previously described protocol (*32*). Briefly, NPCs were single-cell seeded in Virus-NEM Cocktail onto laminin-coated (Sigma) plates at 200,000 cells/cm^2 on Day 0. Virus-NEM Cocktail was prepared as 7 ml NEM, 1 ml rtTA virus, 1 ml ASCL1-puro virus, 1 ml DLX2-hygro virus, and 5 µM ROCK inhibitor. On day 1, we withdrew virus and fed cells with fresh NEM supplemented with 2 µg/ml doxycycline (Sigma). From Day 2 to Day 6, cells were treated in NEM supplemented with 2 µg/ml doxycycline, 2 µg/ml puromycin (Thermofisher), and 100 µg/ml hygromycin (Thermofisher) to remove non-transduced cells. On day 7, we switched media to conditioned NeuralbasalPlus Media and changed media every 3 days. NeuralbasalPlus Media was conditioned in cultured rat astrocyte (Thermofisher) for 24 hrs before media change, and BDNF, GDNF, NT-3, doxycycline were added freshly. Doxycycline was withdrawn on Day 16 and 50 nM Ara-C could be included in media if non-neuron cells were observed. On Day 28, neurons were harvested using Accutase for ATAC-seq or RNA isolation.

1.5. Differentiation into dopaminergic neurons

The protocol for the differentiation of dopaminergic neurons (iN-DA) was adapted from Gonzalez *et al. (33).* Briefly, NPCs were seeded as single cells in NEM onto 6 well plates coated with Geltrex (Thermofisher). Dopaminergic priming media was added to the cells at 50% confluence on Day 0. Dopaminergic priming media was prepared by adding to Neurobasal medium (Thermofisher) the following components: N2 supplement (Thermofisher), B27 supplement (Thermofisher), 100 µg/ml Primocin (Invivogen), Glutamax (Thermofisher), 2 µM purmorphamine (Tocris), 100 ng/ml FGF8 (Thermofisher), and 0.25 µM SAG (R&D). Purmorphamine, FGF8, and SAG were added freshly upon use. Media was changed on Days 2, 4, and 6. On Day 7, the cells were replated onto 6 well plates coated with Geltrex at 5x10^5 cells/well and switched to dopaminergic differentiation media. Dopaminergic differentiation media was

prepared by adding to Neurobasal medium the following components: B27 supplement, 100 µg/ml Primocin, Glutamax, 0.2 mM ascorbic acid (Sigma), 10 ng/µl BDNF (PeproTech), 10 ng/µl GDNF (PeproTech), 1 ng/µl TGFβ3 (PeproTech), and 0.1 mM cAMP (EMD). BDNF, GDNF, TGFβ3, and cAMP were added upon use. Media was changed every other day. On Day 30, dopaminergic neurons were harvested using Accutase (Thermofisher) for ATAC-seq and RNA isolation.

1.6. Immunocytochemistry to characterize iPSCs and neurons

For characterization of iPSC and neurons, cells were fixed in 4% PFA (Sigma) for 15 min at room temperature. After three brief washes in PBS, cells were permeabilized with 1% Triton X-100 in PBS for 15 min at room temperature and further blocked with 3% BSA and 0.1% Triton X-100 in PBS. Samples were subsequently incubated using primary antibodies diluted in blocking buffer at 4°C overnight, followed by three PBS washes. The samples were then incubated with secondary antibodies at room temperature for 1 hr. Both primary and secondary antibodies were diluted in PBS containing 3% BSA. Neurons from the same differentiation experiment were fixed and stained with the same antibody diluent preparation concurrently. After another three PBS washes, samples were incubated in 2.5 µg/ml DAPI (4', 6-diamidino-2-phenylindole) at room temperature for 10 min and mounted on glass slides using Invitrogen Antifade Gold mounting reagent. The images were taken by a Nikon ECLIPSE C2 confocal microscope. Primary antibodies used: anti-SSEA-4 (Abcam, 1:100), anti-NANOG (Abcam, 1:200), anti-Oct4 (Abcam, 1:200), anti-TRA-1-60 (Abcam, 1:300), anti-TUJ1 (EMD Millipore, 1:200), anti-NESTIN (Abcam, 1:300), anti-PAX6 (BioLegend, 1:300), anti-SOX2 (Abcam, 1:1000), anti-MAP2 (Synaptic System, 1:700), anti-GABA (Sigma-Aldrich, 1:1000), Anti-nuclei antibody (Sigma MAB1281, 1:100), anti-VGAT (Synaptic Systems, 1:1000), and anti-vGlut1 (Synaptic Systems, 1:100), Secondary antibodies: Alexa 488 donkey anti-mouse (1:1000), Alexa 594 donkey anti-mouse (1:1000), and Alexa 488 donkey anti-rabbit (1:1000), Alexa 594 donkey anti-rabbit (1:1000).

2.1. RNA preparation, quantitative real-time PCR (qPCR) and bulk RNA-seq

For cell cultures that were subject to gene expression analyses by qPCR or bulk RNA-seq, the culture media was removed by aspiration and the cells were washed twice using 1 ml PBS. Subsequently, 600 µl of RLT Plus reagent was directly added into each well to homogenize cells *in situ*. Qiagen RNEasy plus kit was used for the extraction of total RNA according to the manufacturer's protocol. Extracted RNAs were eluted into 50 µl of RNA-free water, and the

concentration was determined using a NanoDrop-8000 spectrophotometer. For qPCR, reverse-transcription was performed using the ThermoFisher High-capacity RNA-to-cDNA reverse transcription kit with random hexamers according to the manufacturer's protocol. Briefly, 200-500 ng of total RNA was used for each 20 µl RT reaction. The reaction products were then diluted with 80 µl of RNase-free water for qPCR analysis. qPCR was performed on a Roche 480 II instrument, using gene-specific FAM-labelled TaqMan probes (ThermoFisher) for detecting gene expression, with *GAPDH* as the internal control. 3-4 biological replicates (different cultures) were included for each experimental group, and 3 technical replicates were included in qPCR. For bulk RNA-seq, total RNAs extracted from each sample, passing the standard QC metrics (RNA quantity >2 µg, RIN >9.0, $OD_{280/260}$ >1.8), were sent to Novogene for library preparation and sequencing.

2.2. Bulk RNA-seq data processing and differential gene expression (DGE) analysis

RNA-seq files were provided in the format of 2×150 bp paired-end (~25 M reads per sample) fastq files from Novogene. Briefly, libraries were prepared using the NEB Nextera kit with customized adapters. Since the sequencing facility had performed pre-cleaning on raw reads, no Trimmomatic adapter-trimming was performed. 20~30 M reads were recovered from each sample. Raw files were subsequently mapped to human hg38 genome (GRCh38p7) using STAR v2.6.0

(`--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts --outSAMattrIHstart 0 --outSAMstrandField intronMotif --outSAMmultNmax 1 --outFilterIntronMotifs RemoveNoncanonical --outBAMcompression 10 --outBAMsortingThreadN 20 --outBAMsortingBinsN 20 --outFilterMultimapNmax 1 --outFilterMismatchNmax 1 --outSJfilterReads Unique --limitBAMsortRAM 10000000000 --alignSoftClipAtReferenceEnds No --quantTranscriptomeBAMcompression 10 10`).

PCR and optical duplicates were further removed using `Picard MarkDuplicates`, and unique mappers (MAPQ = 255) were retained for downstream analysis. GENCODE v18-based gene annotations, transcript length, and GC percentage were obtained from ENSEMBLE bioMart. For gene-based quantifications, the number of fragments at the meta-feature (gene) level was counted using the `featureCounts` module of R package `SubRead` (*34*), and only genes expressed in at least one cell type with counts per million (CPM) >1 in at least half of the sample within the group were retained for subsequent analysis. Quantile-normalization was

applied to all samples using gene length, GC %, and read depth to calibrate bias between samples using R package `CQN` (*35*). Normalized gene-level FPKM counts were used to plot the correlation heatmap and PCA analysis. The correlation heatmap and PCA map were generated using R packages `gplots` and `ggplot2`, respectively. DGE analysis was performed using `edgeR` by fitting the dataset into a gene-wise negative binomial generalized linear model with quasi-likelihood as `glmQLFit()` and evaluated by `glmQLFTest()`. For transcript-based quantifications, the pseudo-alignment probabilistic model of Kallisto was applied (*36*).

2.3. Integrative analysis of bulk RNA-seq for principal component analysis (PCA)

PCA was used for clustering different cell types differentiated from iPSCs and for comparing to other brain cell types based on their genomic features. In addition to the samples used in the current study, the expression profiles of tissue/cell types from publicly available databases were acquired from GTEx (GTEx Brain, www.gtexportal.org), BrainSpan (prenatal and postnatal sets, www.brainspan.org), and GEO106589 (*37-39*). The gene-level count of all datasets was unified by their stable ENSEMBL gene identifier, and only genes used during RNA-seq analysis (*i.e.*, genes that exhibited expression in at least one cell type with counts per million (CPM) >1 in at least half of the sample within the corresponding group) were used. In order to perform the PCA analysis and correlation analysis, all expression values were converted to $\log_2$RPKM and further quantile normalized using the `normalizeBetweenArrays()` function in the `limma` package of R.

2.4. RNA-seq-based eSNP-Karyotyping for detecting chromosomal aberrations of iPSCs

In order to identify any potential chromosomal aberrations (including chromosomal duplications, loss of heterozygosity, and meiotic recombination) in the iPS cell lines and neuronal cultures, the e-Karyotyping analysis was performed as described (*29*) using the available RNA-seq data. RNA-seq data in paired-end fastq format were aligned to GRCh38.p7 genome using STAR, and the common SNP list from dbSNP 150 was used to generate the moving median plots and heterogeneity maps (Fig. S4B-C) using code adapted from github.io/BenvenLab/eSNPKaryotyping with modifications and GATK 4 integration (*29*).

2.5. Sample identity confirmation using VerifyBamID

As part of the sample quality control (QC), we verified the sample identify of all the iPSC lines and their derived neural cultures using RNA-seq and ATAC-seq data. Briefly, we started with an imputed genotype format that contained around 5.5 million SNPs imputed from an original set of genotypes of ~670,000 autosomal SNPs for the MGS cohort (*9*). Using plink (v1.9), we generated corresponding VCF files for all twenty individuals in each of the iN-Glut and NPC samples, as well as the "core-8" individuals in each of the iN-DN, iN-GA, and iPS cells groups. Next, we ran the liftOver (*40*) tool from UCSC to remap all SNPs in the VCF file to the hg38 reference genome. We used VerifyBamID (v1.1.3) (*41*) to verify whether the reads in a specific BAM file (ATAC-seq or RNA-seq) matched the previously known genotypes of the same sample. We matched samples by comparing their identity by descent (IBD) values generated by VerifyBamID. We found that the IBD values were greater than 0.95 for all the correctly best-matching sample of all the individuals. For the twenty individuals in the iN-Glut-20 group, the mean and standard deviation of IBD values were 0.971 and 0.004, respectively.

3.1. ATAC-seq sample preparation

ATAC-seq samples were prepared as described in (*8*). Briefly, cell cultures were dissociated by Accutase and counted. About 75,000 viable cells were collected and lysed in resuspension buffer containing 0.1% NP40, 0.1% Tween, and 0.01% Digitonin. Nuclei were then collected through centrifugation and resuspended in transposition (with Tn5 transpose) mixture. After a 30 min reaction at 37°C, DNAs were collected using Zymo DNA Clean and Concentrator-5 Kit (D4014). The eluted DNAs were used for for library preparation and ATAC-seq, following the standard protocol, at the University of Minnesota Genomics Center.

3.2. ATAC-seq data pre-processing and quality control (QC)

All raw sequence reads generated by Illumina HiSeq 2000 had been demultiplexed at the University of Minnesota Genomics Center and provided as 2×50 bp paired-end fastq files (targeting 100 M reads per sample, actual reads per sample were between 50~100 M). Adapter remnants, low-quality reads, and low QSEQ short sequences near either end of reads were processed by Trimmomatic (`ILLUMINACLIP:NexteraPE-PE.fa:2:30:7, SLIDINGWINDOW:3:18, MINLENGTH:26`). The processed sequences were separated into paired-end and single-end fastq files per sample, respectively. The paired-end and single-end fastq

files were individually mapped against the human genome reference file including decoy sequences (GRCh38p7/hg38, 1000 Genome Project) using bowtie2 (`-x 2000, -mm --qc-filter –met 1 –sensitive –no-mixed -t`) and subsequently merged and sorted as BAM-formatted files using samtools, with only uniquely mapped reads (MAPQ >30) retained. Picard tools `MarkDuplicate` was then used to remove all PCR and optical duplicated reads from the BAM file. The insert size distribution histograms of each sample were individually generated using Picard `CollectInsertSizeMetrics`. All analyzed ATAC-seq samples passed standard QC based on the characteristic nucleosomal periodicity of the insert fragment size distribution and high signal-to-noise ratio around transcription start sites (TSS) (Fig. S5).

3.3. ATAC-seq OCR peak calling and QC

In order to increase sample size and sensitivity for peak detection, the BAM files of the processed reads within each cell type (8 samples per cell type, thereby called the core-8 group) were merged using Samtools 1.8 before feeding to the peak-calling pipeline (*42*). MACS2 was used to generate peak files (narrowPeak format) of each cell type with recommended settings at FDR = 0.05 (*4*) (`-f BAMPE, --nomodel, --call-summits –keep-dup-all -B`). Peaks that fell within the ENCODE blacklisted regions were removed. Also, we removed peaks falling within chromosomes X and Y, and the mitochondrial genome region. In all, from the core-8 cell lines, we obtained 256,947 peaks from iN-Glut neurons, 281,210 peaks from iN-DN neurons, 278,416 peaks from iN-GA neurons, 256,085 peaks from neural progenitor cells, and 336,835 peaks from iPS cells.

To increase the peak resolution and for comparing with other chromatin peak datasets, we re-centered peaks at their summits and restricted the size of peaks to 501 bp (± 250 bp of the summit). In order to generate a set of non-overlapping, linear peaks representing reproducible peaks from all five cell types that could be used for cross-type comparisons as reported in (*3*), we generated a cumulative peak set by first combining all five peak sets obtained from different cell types without merging overlapping intervals. Subsequently, we only kept the most significant peak by comparing their maximum -log *P* value (peak score) within any peaks that were found to be overlapping, thereby generating a set of non-overlapping linear peaks (n = 666,614) that reflected the chromatin states in all five cell types. Such peak intervals were used as the basis in generating the PCA plot and for hierarchical clustering. When generating density-dependent scatter plots, we used a

random subset of 20,000 peaks from the shared peaks within the compared samples to achieve sufficient resolution (Fig. S6B) (*3*).

For QC of OCR peak calling, we confirmed that peak intensities (normalized ATAC-seq read counts) were highly correlated between samples within a cell type (R > 0.9) (Fig. S6B). We confirmed the OCRs for some known cell-type-specific genes, such as *GAD2*, encoding an enzyme responsible for the synthesis of GABA in iN-GA, and *NANOG*, encoding a transcription factor (TF) essential for maintaining pluripotency in iPSCs (Fig. S6D, E). We then clustered 40 ATAC-seq samples (8 samples each × 5 cell types) using quantile-normalized reads of a common set of 666,614 non-overlapping peaks. Using the same set of OCR peak intervals, we further carried out PCA to compare our ATAC-seq data with publicly available open chromatin datasets from fetal cortical organoids (*12*), fetal brains (*11*), and PsychENCODE adult brains (*3*).

3.4. PCA of ATAC-seq and DNaseI HS-seq datasets

This is to compare the similarities of chromatin accessibility between our ATAC-seq results in different cell types and previously assayed adult/fetal brain cells. We acquired chromatin accessibility data from ATAC-seq of post-mortem brain (*3*), brain organoid (*12*), as well as DNase I hypersensitive assay from the fetal brain (NIH Roadmap) (*11*). Subsequently, we applied our above described cross-cell-type non-overlapping OCR peak interval (n = 666,614) (GTF file, genome lift-over was performed when necessary) to the acquired raw data using `featureCounts` (*34*) to get per-interval read count. PCA was applied to the assembled datasets, including both our core 5 cell types and the output from these public datasets.

3.5. Comparison of neuronal ATAC-seq peaks with PsychENCODE brain OCR peaks

To compare our peak sets to the PsychENCODE adult brain ATAC-seq peaks (*3*) we re-centered the PsychENCODE peaks to 501 bp in length (± 250 bp of the summit, 117,935 peaks in total). The initial comparison used our peak sets called under the FDR = 0.05 (`-f BAMPE, --nomodel, --call-summits –keep-dup-all -B`) (see section 3.3). In order to make our peak sets comparable (*i.e.*, parameters used for peak calling and the total number of peaks in both datasets) to PsychENCODE peaks, we used our peaks with FDR cut off of 0.01 instead of 0.05 to match PsychENCODE peaks. We also re-centered our peak set to 501 bp in length as ± 250 bp of the summit as done for PsychENCODE peaks. We defined the shared peaks as with at least 25% overlapping reciprocally (125 bp) in both our peak sets and the PsychENCODE peaks.

4.1. Allele-specific open chromatin (ASoC) SNP calling

GATK (version 4.0) was used for SNP calling (*43*) for allelic imbalance of chromatin accessibility, as recommended by the GATK Best Practices (`software.broadinstitute.org/gatk/best-practices/`). Briefly, to account for mapping bias to reference alleles in allele-specific analysis, WASP-calibrated BAM files generated from ATAC-seq pipeline were used as input (*44*). Variants were called using the discovery mode of `HaplotypeCaller` with human GRCh38 (hg38) genome and the corresponding dbSNP version 150, and only reads with MAPQ score ≥30 were used (`-stand_call_conf 30`). Subsequently, recalibration of SNPs and Indels were performed in tandem using the `VariantRecalibrator` function (`-an DP -an QD -an FS -an SOR -an MQ -an ReadPosRankSum -mode SNP -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 -mG 4`) and applied using `ApplyRecalibration`. Databases used for `VariantRecalibrator` include the hg38 versions of HapMap v3.3 (`priority = 15`), 1000G_omni v2.5 (priority = 12), Broad Institute 1000G high confidence SNP list phase 1 (`priority = 10`), Mills 1000G golden standard INDEL list (`priority = 12`), and dbSNP v150 (`priority = 2`). Each sample within the cell type was processed individually, and heterozygous SNP sites whose tranche level >99.9% were extracted. To reduce bias introduced by any acquired (or "*de novo*") mutations during cell growth, only SNPs with corresponding rs# records found in dbSNP v150 were retained. For each of the five cell types, all heterozygous sites that passed the filter above (8 samples for each cell type) were merged by `CombineVariants` to produce the master VCF file of the cell type (Fig. S8). For iN-Glut and NPC sets, VCF files derived from their 20 samples were also generated separately.

To maximize the power to detect ASoC, we pooled reads for all the called heterozygous SNPs from the core-8 or the extended 20 iN-Glut/NPC samples as justified by the high concordance of allele-specific effects across individuals (see the section below "Testing the inter-individual concordance of the directionality of allelic chromatin accessibility"). Finally, the VCF files were filtered and only biallelic SNP sites (GT: 0/1) with minimum read depth count (DP) ≥20 and minimum reference or alternative allele count ≥2 were retained. The binomial *p*-values (non-hypergolic) were calculated using the `binom.test(x, n, p = 0.5, alternative = "two.sided", conf.level = 0.95)` from the R package, and FDR correction were applied using the number of all qualified SNPs as the correcting factor of R function `p.adjust(x,`

`method = "fdr"`). We set the threshold of ASoC SNP at FDR-corrected binomial *p*-value = 0.05.

To estimate the ASoC SNP detection sensitivity at a given OCR read-depth, we plot the percentages of heterozygous SNPs that were detected as ASoC SNPs (FDR < 0.05) in iN-Glut-20 dataset at different cut-offs of OCR read-depth at the SNP sites (*e.g.*, >20, >100, >200) (Fig. S9D). At each cut-off value, only the heterozygous SNPs with read depth above that cutoff was evaluated. To estimate how the ASoC detection sensitivity varies with SNP effect sizes at give-depth, we plot the percentages of all heterozygous SNPs with allelic ratio above a certain cut-off that were detected as ASoC SNPs in iN-Glut-20 dataset (FDR < 0.05) at three different cut-offs of OCR read-depth (>20, >100 and >200) (Fig. S9E).

The correlation map of allelic ratios between different cells types was generated using the `cor.test(alternative= "two.sided", method = "pearson")` function of stats package in R, and results were visualized by `heatmap.2()`. Read pileup proximal to the SNP site was generated using Samtools mpileup function, and differentiation of allele-specific reads was performed using the `SNPsplit` Perl package (www.bioinformatics.babraham.ac.uk/projects/SNPsplit/). The final readouts from both read pileup and SNP-specific reads were visualized using the R package `Gviz`. We confirmed no obvious mapping bias to reference alleles by visualizing the volcano plots that graph the allelic read-depth ratios against -log2 *P* values in scatter plots (Fig. S9C).

4.2. Testing inter-individual concordance of the directionality of allelic imbalance

This analysis serves to justify the approach of pooling all the ATAC-seq samples of a cell type to increase the power of detecting ASoC SNPs (*8, 13*). In order to compare whether the same SNP site often shows concordant directionality of allelic bias between individuals, we analyzed all 20 iN-Glut ATAC-seq samples. We first called all heterozygous, biallelic SNPs within each cell line, and identified ASoC SNPs with a binomial test as described above, requiring a minimum read depth of 20 and the minimum number of reference and alternative alleles as 2, *i.e.*, the same metrics used for calling ASoC in the pooled samples. Since calling ASoC in any single individual has limited power, we relaxed the statistical significance cutoff to P < 0.05 to increase the possible "ASoC" SNP pairs between individuals. We then calculated the fraction of SNP allelic effect directionality by comparing the shared ASoC SNPs in each pair of the iN-Glut-20 group. Briefly,

heterozygous SNPs shared in two samples were classified as either having allelic bias (*i.e.*, ASoC in both samples) or not (not ASoC in either sample) and compared separately. This resulted in 40,020 pairs for SNPs assigned as having allelic bias and 1,588,594 SNP pairs as not having allelic bias, and then the concordance of the directionality of allelic chromatin accessibility was compared (Fig. S9A).

4.3. Testing association of allelic imbalance with known sample traits

In order to investigate whether any of the properties of the 20 individuals used in the study could have impacts on allelic imbalance, we applied a linear model using the three known traits (sex, age, and SZ status) against the ratio of reference allele. None of the three known traits was found to affect ASoC (*i.e.*, all with FDR > 0.5).

4.4. Cell type specificity analysis of ASoC

We used Storey's π1 analysis (*45*) to evaluate the pairwise sharing of ASoC SNPs among all 5 cell types (Fig. 1D). We have chosen Storey's π1 analysis (*45*) because it accounts for the incomplete power of calling ASoC in the matched cell type, and is often used to investigate sharing of eQTL between cell types (*46*). For each pair of 'leading' and 'matched' cell type, we obtained the ASoC SNPs (FDR < 0.05) from the leading cell type and estimated the proportion of non-null tests (π1) based on the binomial *p* values of these SNPs in the matched cell type. Note that if a SNP had a DP < 20 in the matched cell type, it was automatically counted toward the null proportion.

To define neuron-specific ASoC in the comparison of chromatin accessibility between two cell types (Fig. 1E and Fig. S11), we require a SNP to be ASoC at FDR < 0.05 in one of the neuronal cell types, but read depth < 20 or binomial test of ASoC p > 0.05 in iPSC.

5.1. Enrichment of ASoC SNPs in annotated genomic/epigenomic elements

In this analysis (Fig. 2A and Fig. S12), we define cell-type-specific ASoC SNP as those with FDR < 0.05 in one cell type, and their nominal *p*-values > 0.05 in all other cell types. To define cell-type-shared SNPs, we require a SNP to have FDR < 0.05 in at least 3 cell types.

We examined whether ASoC SNPs are enriched in regulatory DNA sequences marked with the annotated genomics/epigenomics features. This analysis is similar to that implemented in GREAT and was described previously (*4, 47, 48*) for testing the folds of peak enrichment within

the annotated genomic regions and epigenetically annotated regions. The enrichments were calculated using the formula below:

$$\text{ratio} = \frac{\left(\frac{\text{Number of bases within the peaks of cell type that overlap feature}}{\text{genome size}}\right)}{\left(\frac{\text{Number of bases within feature}}{\text{genome size}}\right)\bullet\left(\frac{\text{Number of bases within the peaks of cell type}}{\text{genome size}}\right)}$$

For ASoC SNPs that were cell-type-specific or shared by three cell types, we analyzed enrichment considering SNP intervals as the unit similar to OCR peaks. The enrichment *p*-values were estimated based on the binomial test as implemented in GREAT and as previously described (*4, 48*).

$$p_{\text{enrichment}} = \sum_{i=k_\pi}^{n} \binom{n}{i} p_\pi^i \left(1-p_\pi\right)^{n-i}$$

from which we can calculate the p-value using the

`binom.test(x, n, p, alternative="greater")` function in R, where

x = numbers of SNPs that fall within the designated epigenetically annotated features;

n = total length of the SNP intervals (in bp);

p = total length of the designated epigenetically annotated features (in bp) / genome size (3.2×10^9 bp)

The Gene-based annotation of the genome was derived from GENCODE v28 as part of the built-in database of the HOMER package (*49*). The list of human forebrain/non-forebrain enhancers was from the VISTA enhancer browser (*50*). Human-gained enhancers were acquired from GSE63648 using data from either frontal or occipital cortex at 12 PCW and marked by H3K27ac or H3K4Me2 (*51*). The definitions of chromatin state were assembled using an imputed 25-state model derived from individual #E081 of fetal brain tissue by the Roadmap Epigenomics Project (*48, 52*). For estimating the enrichment of ASoC SNPs (Fig. 2A), the categories of epigenomic features used for promoter and enhancer annotations are: promoter = TssA, PromU, PromD1 and PromD2; enhancers = TxReg, TxEnh5, TxEnh3, TxEnhW, EnhA1, EnhA2, EnhAF, EnhW1, EnhW2, EnhAc, and DNase. The abbreviation of different chromatin states is promulgated below, as used in (*48*):

TssA	Active TSS
PromU	Promoter Upstream TSS
PromD1	Promoter Downstream TSS with DNase
PromD2	Promoter Downstream TSS
Tx5'	Transcription 5'
Tx	Transcription
Tx3'	Transcription 3'
TxWk	Weak transcription
TxReg	Transcription Regulatory
TxEnh5'	Transcription 5' Enhancer
TxEnh3'	Transcription 3' Enhancer
TxEnhW	Transcription Weak Enhancer
EnhA1	Active Enhancer 1
EnhA2	Active Enhancer 2
EnhAF	Active Enhancer Flank
EnhW1	Weak Enhancer 1
EnhW2	Weak Enhancer 2
EnhAc	Enhancer Acetylation Only
Dnase	DNase only
ZNF/Rpts	ZNF genes & repeats
Het	Heterochromatin
PromP	Poised Promoter
PromBiv	Bivalent Promoter
ReprPC	Repressed PolyComb
Quies	Quiescent
Forebrain_enh	Forebrain enhancers
Non_forebrain_enh	Non-forebrain enhancers
H3K27acF	H3K27 acetylated regions found in the frontal lobe
H3K27acO	H3K27 acetylated regions found in the occipital lobe
H3K4me2F	H3K4me2 regions found in the frontal lobe
H3K4me2O	H3K4me2 regions found in the occipital lobe

5.2. Reporter gene assay of ASoC SNP effect on promoter/enhancer activity

To examine whether ASoC SNPs are functional and affect the potential promoter/enhancer activity of the short DNA sequence segment that flank the ASoC SNPs, we first performed a data-mining of a massive parallel reporter assay (MPRA) dataset that tested the allelic effects of thousands of SNPs on reporter gene expression in a non-neuronal lymphoblastoid cell line (LCL) (*15*). In a reporter gene assay system, differential promoter/enhancer activity of the synthesized

DNA fragment containing different alleles of a tested SNP was measured by the SNP allelic effect on reporter gene expression. We intersected our lists of ASoC SNPs (iN-Glut-20 and NPC-20) (using a relaxed FDR < 0.1 to increase the number of ASoC SNPs for comparison) with the list of SNPs that altered reporter gene expression in MPRA dataset. For the overlapping SNPs (ASoC vs. MPRA), we estimated whether our ASoC SNPs were enriched for SNPs that also showed differential allelic effects on reporter gene expression in the MPRA. Fisher's exact test was used to estimate the statistical significance of the enrichment. Furthermore, for the overlapping SNPs, we also calculated the Pearson's correlation of ASoC SNP allelic ratio and the differential SNP allelic effect on reporter gene expression.

We also carried out our own reporter gene assay for a list of selected ASoC SNPs to evaluate the allelic effects of ASoC SNPs. The functional read out of this assay was the allelic effect on reporter gene expression. We randomly selected 8 ASoC SNPs (FDR < 0.05 in iN-Glut-20 dataset) associated with SZ (Table S19) for validation in an iPSC-derived NPC line infected by a customized lentivirus-based luciferase reporter construct. pLenti-PGK-V5-LUC-Neo plasmid (Addgene# 21471) was digested with XhoI and SalI to remove PGK promoter 5' upstream of the reporter luciferase (*LUC*) gene. For ASoC fragment cloning, 81 bp sequence with each allele spanning the SNP site (40 bp upstream and 40 bp downstream of the SNP site) together with 32-bp minimal promoter sequence was synthesized as a single-strand oligo and inserted into the linearized vector by Gibson assembly (new construct was named as pLenti-minP-PGK-V5-LUC-Neo). Single bacterial clones were picked and confirmed by Sanger sequencing after transformation. Virus particles were prepared as previously described (*53*). For infection of NPCs, $3x10^5$ cells/well were seeded in 12-well plates with 1ml neural maintenance media containing 100 µl virus supernatant and ROCK inhibitor on Day 0. On Day 1, cells were fed with fresh maintenance media containing 10 ng/µl BDNF, 10 ng/µl GDNF, 10 ng/µl NT-3, and 500 µg/ml G418. Media were refreshed every day. On Day 11, RNA was isolated and *LUC* expression was quantified by qPCR and normalized by neomycin expression. The oligonucleotide sequences flanking each SNP site are available in Table S14. Differential expression of the reporter gene between the two alleles for each assayed ASoC SNP was tested by Student's *t*-test with Welch's correction (Fig. S13F).

5.3. Intersecting ASoC SNPs with brain eQTLs, Hi-C, and promoter annotations

This is to examine whether ASoC SNPs can be mapped to putative *cis*-target genes by interesting with brain eQTLs and Hi-C chromatin contact data as well as promoter annotations. For each ASoC SNP in either iN-Glut-20 or NPC-20 samples, we identified its putative *cis*-target gene(s) using a combination of gene annotations, brain eQTLs, and Hi-C data. If a SNP is located within a promoter, defined as upstream 2kb and downstream 1kb of TSS (GENCODE v31 comprehensive annotation), we assigned the SNP to the gene of that promoter (promoter target). For brain eQTL, we used a high-powered postmortem brain eQTL dataset from the CommonMind Consortium (CMC) (*17*). A SNP was assigned to a gene (eQTL target) if the TSS of that gene is within 500 kb of the SNP, and the SNP is associated with gene expression at FDR < 0.05. This analysis may significantly underestimate the percent of ASoC SNPs that are also eQTLs because our eQTL data is from postmortem/adult brain. To better match the tissue type of ASoC and eQTL data, we performed additional analysis by limiting the ASoC SNPs to those in ATAC-seq peaks overlapping adult brain OCRs from PsychENCODE. Finally, we assigned a SNP to its putative target based on physical chromatin interactions (Hi-C target), measured by promoter capture Hi-C (*18*). We considered all called interactions from excitatory neurons and hippocampal neurons from Table S2 of (*18*). To assign the Hi-C target of an ASoC SNP, we first found the ATAC-seq peak associated with that SNP. We then expanded an interacting Hi-C fragment to a minimum width of 5 kb if the fragment size is less than 5kb as described in (*18*). If an ATAC-seq peak overlapped (at least 20% of peak length) with a fragment in a Hi-C interaction, we extracted the other interacting Hi-C fragment and assigned it to a gene if the fragment overlapped with the promoter of that gene (-2kb/+1kb of TSS).

5.4. Transcription factor (TF)-binding footprint analysis from ATAC-seq data

TF footprints were predicted based on ATAC-seq reads pooled from samples of the same cell type (8 samples for each of the 5 cell types; and 20 samples per cell type in the case of iN-Glut-20 and NPC-20). First, the R package `DNase2TF` (*54*) was used to identify potential TF binding windows of size 6-30 bp within the ATAC-seq peak regions by detecting local read cut count depletion with a Z-score threshold of -2 and an FDR level of 0.05. Next, a set of 522 *Homo sapiens* TF motifs available in the JASPAR database 2018 release were matched against ATAC-seq peak regions using the `rgt-motifanalysis` motif matching function from python library RGT (*55*) with an FPR threshold of 1×10^{-4}. Motif matches that overlapped with the potential TF binding

windows identified by DNase2TF by at least 50% were determined as the final TF binding footprints. To further investigate the averaged ATAC-seq cleavage profiles around ASoC SNPs in/outside predicted TF footprints, we counted the number of 5' ends of ATAC-seq reads in 200 bp windows centered around ASoC SNPs (Fig. S15A).

5.5. TF motif-break analysis: ASoC SNP allelic imbalance and motif disruption score

This analysis was designed to test if a motif/TF may have a role in driving chromatin accessibility. For such a TF, we would expect that genetic perturbation of its binding site would lead to a matched change of chromatin accessibility. For instance, an allele disrupting the motif of a pioneer factor is expected to reduce chromatin accessibility. This analysis is analogous to Mendelian Randomization (*56*) that uses genetic variants to test the causal effect of an exposure (in our case, TF/motif) on the outcome (chromatin accessibility), and has been used in earlier work (*57*).

Specifically, for a given cell type, we first obtained all the ASoC SNPs that are located inside identified TF binding footprints of any of the 522 JASPAR TF motifs. Next, we use the function `motifbreakR()` from R package `motifbreakR` to evaluate the level of disruption the alternative allele of each SNP has on any possible motif that matched its surrounding sequence with a *p*-value filtering threshold of 2.5×10^{-4}. Only SNPs with 'strong' effects of motif disruption (can be either positive or negative) according to `motifbreakR` were kept for further analysis, and their motif disruption scores were quantified as the difference between `scoreAlt()` and `scoreRef()` from `motifbreakR` results. Then, for each given TF motif, we collected all ASoC SNPs that 'strongly' disrupted their TF-binding motifs, and conducted simple linear regression with an offset of zero between motif disruption scores of these SNPs and their allelic imbalance levels. The latter was quantified by taking the $\log_2$ of the ratio between alternative and reference ATAC-seq reads at each of the SNP loci. Polarized $-\log_{10}$ of regression *p*-values were shown on the heatmaps (Fig. 2E and Fig. S16A) for evaluating the significance and direction of the association between motif disruption and allelic imbalance of chromatin accessibility.

Since motifs of related TFs are sometimes similar, we clustered motifs as a post-processing step, and assigned each motif to a cluster number (Fig. S16). We first obtained the RSAT matrix-clustering result on the JASPAR 2020 CORE Vertebrates motifs from the JASPAR database. The JASPAR 2020 CORE Vertebrates motifs were partitioned into 111 clusters, and 355 out of our

522 *Homo sapiens* motifs were included in the clustering result. Next, we use the motif comparison tool Tomtom (*58*) to measure pairwise similarity for all the 522 motifs based on their position weight matrices. Pairs with an E-value < 10 were kept as similar motifs. For each of the *Homo sapiens* motifs that were not included in the JASPAR clustering result, we assign it to the same cluster as the motif it is the most similar to according to Tomtom, and has already been assigned a cluster. In the end, all 522 *Homo sapiens* motifs were assigned into 76 distinct clusters.

5.6. Enrichment analysis of TF-binding motif for ASoC SNPs and OCR using HOMER

We used HOMER (*49*) to assess the enrichment of TF binding motifs in sequences flanking ASoC SNPs (±50 bp) and in OCRs. We used the JASPAR database (2018 release) with all 522 human TF motifs for analysis. For each cell type, we used the ASoC SNPs (FDR < 0.05) positions from the core-8 group ± 50 bp and the set of commonly shared peaks (Fig. S5A) separately as input intervals. The parameters used were `findMotifsGenome.pl <input.bed> hg38 -cpg -mknown jasper2018.known <output>`. The genetic background for the enrichment test was provided by HOMER and the significance of the enrichment were derived from HOMER.

5.7. Enrichment analysis of TF- binding foortprints (or TFBS) for ASoC SNPs

We first identified TFBSs in glutamatergic neurons for all 522 *Homo sapiens* TF motifs available in the JASPAR database (2018 release) by applying the footprint calling method as described above on ATAC-seq reads pooled from iN-Glut-20 as well as for NPC-20 cells as well as from "core-8" samples of each cell type. Then, for each TF motif in each cell type, we evaluated the enrichment level of ASoC SNPs (FDR < 0.05) vs non-ASoC SNPs in OCRs (FDR > 0.05) in all the identified TFBSs of this motif using Fisher's exact test.

6.1. Enrichment analysis of SZ GWAS risk variants using index SNPs

In order to examine whether SZ-associated variants are enriched in our ASoC SNP list, we first compiled a list of SZ GWAS index SNPs (at 108 loci) and all their LD proxies ($r^2 \geq 0.8$; n = 3,507) (*10*). We then intersected the list with the ASoC SNP table (FDR < 0.05) from iN-Glut-20 as well as all heterozygous SNPs found (n = 5,590 and n = 106,030) respectively. The enrichment of ASoC SNPs (vs. non-ASoC SNPs) for SZ GWAS index SNPs was calculated using Fisher's exact test.

6.2. Enrichment analyses of ASoC for brain QTLs and GWAS risk variants using TORUS

We applied a Bayesian hierarchical model (TORUS) to perform a SNP-based enrichment analysis, testing whether risk variants from GWAS are enriched in given functional genomic annotations, *e.g.* OCR and ASoC in our case (*59*). TORUS assumes that every variant is a risk variant or not, represented by a binary indicator variable (1 or 0). The prior probability of the indicator of a SNP being 1 depends on its annotations. TORUS links GWAS effect sizes of SNPs and their annotations by

$$\beta_j = (1-\pi_j)\delta_0 + \pi_j g(\cdot)$$

$$\log\frac{\pi_j}{1-\pi_j} = \alpha_0 + \sum_{k=1}^{m}\alpha_k d_{jk}$$

where GWAS effect size β_j follows a spike-and-slab distribution *a priori*, and π_j is the prior probability for the j-th SNP in a certain locus, modelled by a logistic link with annotation d_{jk} (for the k-th annotation) for SNP j. Usually a normal prior distribution is used for $g(.)$. For binary annotations (1 if a SNP has that annotation, 0 otherwise), the parameter α_k, is the log odds ratio of k-th annotation, and measures enrichment of risk variants in the k-th annotation, relative to all SNPs in the genome that do not have that annotation. TORUS uses the summary statistics of the entire genome to estimate the enrichment parameters.

For brain molecular QTL analysis (Fig. 2B), we used brain xQTL data that included expression, methylation, and histone acetylation QTLs (*16*). For this analysis, we considered ASoC SNPs in iN-Glut, treating ASoC as a binary annotation: 1 if the SNP has ASoC and 0 otherwise.

For GWAS enrichment analysis, we evaluated 17 genomic features: 5 from ATAC-seq peaks of different neuron cell types, 2 from identified allele-specific open chromatin variants, and another 8 from baseline generic genomic features, including coding, intron, promoter, conserved sequence, 3' UTR, 5' UTR, DNase I hypersensitive (DHS) site and repressed regions (*2*). All the annotations are encoded as binary (1 if a SNP has an annotation, *e.g.*, in an ATAC-seq peak, and 0 otherwise). We conducted both univariate and joint analysis in TORUS to assess enrichment of each aforementioned genomic features, as well as their joint contributions. Three combinations of joint analysis were performed: 1) ATAC-seq peaks and allele-specific open chromatin variants; 2) all baseline generic features; 3) all 17 genomic annotations combined. The GWAS datasets used

for enrichment/TORUS analysis were from multiple sources, including both neuropsychiatric disorders and control disorders/traits, as listed in Table S21.

6.3. Fine mapping of putative causal SNP(s) in SZ GWAS

We defined candidate regions for fine-mapping for the 20 selected SZ-associated ASoC SNPs as reported in Table S30 on 108 genome-wide significant loci (genomic regions) for SZ (*10*), and for three SNPs that were not annotated in the Table 30, we used the genome-wide linkage disequilibrium (LD) blocks from the European population (*60*). Specifically, the genome was partitioned into 1,703 roughly independent genomic regions based on European samples from the 1000 Genomes Project.

We used SuSiE to fine map the twenty regions defined above, while incorporating functional information of SNPs as informative priors in SuSiE. We used the summary statistics version of SuSiE (www.biorxiv.org/content/10.1101/501114v1), with the external LD from the 1000 Genomes Project. It is known that fine-mapping using summary statistics can be sensitive to mismatch of external LD and in-sample LD, which is generally not available for large meta-analysis. So, we ran Susie with L = 1, *i.e.*, a maximum number of causal variants in a region is 1. This ensured that the results were not dependent on LD. Our fine-mapping method incorporated multiple genomic features to favor SNPs likely to be functionally important, *e.g.*, ASoC variants. Specifically, we modeled prior inclusion probabilities, as functions of annotations of the SNPs, including ATAC-seq peaks of 5 neuron cell types, two types of allele-specific open chromatin variants and other genomic features, such as coding and promoter. These prior inclusion probabilities were obtained from the results of TORUS using all annotations jointly, as described above. Posterior inclusion probabilities (PIP) were then computed for each SNP using SuSiE. It quantifies the evidence of a SNP being a causal SZ variant with PIP from 0 to 1.

7.1. CROP-seq: overall design of the multiplex CRISPRi/scRNA-seq

To systematically identify possible *cis*-target genes of the regulatory sequences tagged by a set of 20 selected ASoC SNPs that were associated with SZ, we adapted a CROP-seq approach (*20*) with modifications. Briefly, lentivirus expressing guide RNAs (gRNAs) that targeted to specific ASoC sites were used to infect NPCs (form 3 iPSC lines) stably expressing dCas9/Krüppel associated box (KRAB), where KRAB would repress possible *cis*-genes of specific ASoC sites targeted by gRNAs. Combined with single-cell RNA-seq (scRNA-seq), this approach allowed us

to identify the *cis*-target genes of an ASoC sequence by comparing gene expressions between single cells infected by a ASoC-specific gRNA vs. those infected by control negative gRNAs. For each SNP site of interest, we designed three sgRNAs that targeted sequence within ±150 bp of the SNP site on both forward and reverse strands, with at least two sgRNAs target sequences located on the opposite strands at either side of the SNP site. Benchling (benchling.com) was used to generate candidate sequences and only sequences with the highest specific scores (further confirmed using crispr.mit.edu) were used for oligo synthesis by IDT. The selected sgRNA sequences did not overlap with any common heterozygous SNPs as examined from ATAC-seq data. The synthesis, pooling, and expansion of sgRNA-carrying constructs were performed as described below.

7.2. CROP-seq: preparation and packaging of lentivirus particles

The protocol was adapted from the published protocols of Datlinger *et al.* and Xie *et al.* with minor modifications (*20, 61*) We first established a human NPC line with stable expression of the dCas9-KRAB fusion protein by transfecting the cells with human lentivirus particles pre-packed with a pLenti-dCas9-KRAB-Blast construct in 293T cells (*53*). Briefly, 293T cells were maintained in 6-well plates in DMEM media supplied with 10% fetal bovine serum (FBS) and antibiotics at 37°C and 5% CO_2. One day before transfection, 293T cells were seeded at a density of 2 x10^6 per well in T-25 flasks. At the day of transfection, the culture media was replaced fresh with reduced FBS (5%) and without antibiotics. Lenti-dCas9-KRAB-Blast plasmid (Addgene #89567) was co-transfected with pMD2.G (Addgene #12259) and psPAX2 (Addgene #12260) at a 4:2:3 molar ratio using FuGENE HD (Promega) according to vendor's protocol. All the other viral plasmids including rtTA, Ascl1-puro, Dlx2-hygro, CROPseq-Guide-Puro gRNA library were co-transfected with pMDLg/pRRE (Addgene #12251), pMD2.G (Addgene #12259) and pRSV-Rev (Addgene #12253) at a 1:1:1:1 molar ratio using FuGENE HD. 24 hours after transfection, the media was replaced with fresh neural maintenance media (49 ml of Neural Basal Medium, 49 ml Advanced DMEM-F12, 2 ml of Neural Induction Supplement, and 0.2 ml of Primocin). 40 hrs post-transfection, virus-containing media were collected and applied to centrifuge at 500 g for 5 min to spin down debris. The supernatant was aliquoted into 1.5 ml low protein binding tubes and stored at -80°C.

For preparing the sgRNA library, ssDNA oligos carrying sgRNA sequence were pooled in equal amounts and then cloned into CROPseq-Guide-Puro (Addgene #86708) backbone by Gibson Assembly. The assembled constructs were electroporated into Lucigen Endura competent cells. A small aliquot of bacteria was diluted and plated onto 10 cm LB plates for verification of library coverage, and the remaining bacteria were plated onto 25×25 cm LB plates. Twenty hours post replating, bacteria were collected to isolate plasmid using QIAGEN endo-free maxi kit according to vendor's instructions. The colonies on 10 cm LB plates were counted to calculate library coverage and picked to check gRNA representation by Sanger sequencing.

To calculate the virus titer, NPCs were plated at 50,000 cells per well in 24-well plate. 24 hrs post replating, cells were transduced by virus prepared in 1:5 series for 5 dilutions (from 1:5 to 1:3125) using culture media. 48 hrs post-transduction, cells were selected by 1 µg/ml puromycin for 72 hrs. After selection, viable cells in each well were collected using Accutase and counted using a hemocytometer to calculate titer.

7.3. CROP-seq: lentivirus transduction and single-cell CRISPRi perturbation in NPCs

To make NPC lines stably expressing the fused dCas9/KRAB, NPCs were single-cell seeded onto Geltrex-coated 6-well plates at 150,000 cells per cm^2 in 1 ml of Neural Expansion Media containing 0.1 ml of dCas9-KRAB-Blast lentivirus. The virus-containing media were removed 24 hr post-transduction and replaced by selection media (neural maintenance media with 15 µg/ml blasticidin). Blasticidin was included in media through the experiment to prevent from possible silencing of the dCas9-KRAB-Blast construct. The selection media was refreshed daily for the first 72 hrs and subsequently every 48 hrs for the growth and expansion of NPC.

For transduction of dCas9-KRAB-expressing NPCs, the dCas9-KRAB-expressing NPCs were seeded onto Geltrex-coated 6-well plates at a density of 2×10^5 cells per cm^2 one day before transduction. At the day of transduction, the culture was replaced with media containing the pool of CROP-seq sgRNA virus. To maximize the number of single cells that were infected by single gRNA, the added sgRNA library virus was calculated at a low titer (MOI=0.2). 24 hrs post-transduction, the virus-containing media were replaced with fresh media. Drug selection was performed by replacing the culture media containing 1 µg/µl of puromycin 48 hrs post-infection. The NPCs were selected for 72 hrs before harvesting, whereas culture media was replaced fresh

daily. Surviving cells were harvested using Accutase and resuspended with Neural Expansion Media with 10% DMSO. The collected cells were frozen at -80°C until scRNA-seq.

The collected NPC samples were processed with 10x Genomics Chromium Single Cell 3' Reagent Kits v3 according to the manufacturer's instructions. Sequencing of the prepared library was performed on an Illumina HiSeq 2000 sequencer generating approximately 450 M reads in total (at University of Minnesota Genomics Center).

7.4. CROP-seq: single-cell RNA-seq data processing

10x Genomics Cellranger v2.1.2 was used to process raw sequencing data. Briefly, sequencing data were aligned to the human GRCh38/hg38 genome with spike-in gRNA sequences as artificial chromosomes (20 bp gRNA sequence and 250 bp downstream plasmid backbone per each gRNA). The full sequence of CROP-seq-Guide-Puro construct was also included as a decoy to capture vector-borne sequences. The alignment was performed using STAR 2.5.1 and a customized spiked-in version of GENCODE v28 GTF file containing the aforementioned gRNA sequences and vector backbone as annotations. Gene count matrix (in HDF5 format) were used in subsequent R analysis. In all, we achieved a total mapping rate of 94.1%, with 74.4% mapping rate to the human transcriptome.

The `cellrangerR` package (2.1) was used to extract information and assemble data tables from the HDF5 files to generate the digital gene expression matrix, which was constructed based on per-gene UMI count. Cells with more than 500 genes and 11,000 UMIs detected were retained (4,099 out of 4,144 cells) for subsequent analysis. For each cell, the mean read count was 109,945, the median of UMI was 28,480, and the median of genes per cell was 4,935. A cell assigned with a unique RNA was defined as the UMI count of its dominant gRNA in the cell was at least three times more than the sum of UMI counts from all other gRNAs. 2,522 cells (60.8%) survived the filter and designated as cells with uniquely assigned gRNA for further study. T-SNE analysis was performed using the in-built function of the `cellrangerR` package and subsequently plotted using R package `ggplot2`.

7.5. CROP-seq: screening for *cis*-targets of ASoC sequences

We used CROP-seq as a screening tool to identify cis-target genes of the 20 selected ASoC SNP-flanking sequences. The cis-targets were defined as those that showed differential expression

(DE) between the group of single cells assigned with an unique gRNA and the cells infected by the negative control gRNAs. DE analysis in single cells was performed using the `edgeR` package by applying a general linear model (glm) to the dataset, and by identifying DE genes using the `edgeRQLFDetRate` test (*62*). We selected this model because it had been shown to perform better than most other methods in identifying differentially expressed genes in scRNA-seq experiments with small cell numbers (range of 6-400 cells) (*62*). While each sequence was targeted by 3 gRNAs, we found that the efficiencies of gRNAs often differed from each other (as measured by the change in expression of their targeting genes), so we analyzed the cells designated to each gRNA individually. Cells containing the negative control gRNAs (3 target EGFP and another 2 target scrambled sequences) were used as the control group in DE analysis. For DE analysis, we only considered the putative *cis*-gene within 500 kb of each targeted SNP. Only genes with CPM >30 in more than 20% of cells were kept for DE analysis. For DE analysis for each gRNA, to correct for *p*-value inflation commonly seen in scRNA-seq, we performed permutations by permuting the label of cells (cases or controls) 30 times. We then used *p*-values from all genes included in the permutations as the null distribution. We then used this null distribution to assign an empirical *p*-value of each gene, as the number of genes with *p*-values less than or equal to the observed test statistic in the null distribution divided by the total number of genes in the null distribution. Please note that the p-values reported in Fig. 4B and Fig. S22 are nominal p-values (calibrated by permutation), but not corrected for multiple testing.

7.6. Independent CRISPRi/qPCR validation of the *cis*-genes identified from CROP-seq

Because we considered CROP-seq as a tool of screening for a *cis*-target of a putatively functional DNA sequence tagging the ASoC SNPs, we carried out independent validation of a selected set of ASoC *cis*-target genes nominated by CROP-seq. We performed a multiplex CRISPRi in NPCs, which was followed by qPCR validation of the *cis*-targets. We have selected 9 CROP-seq cis-targets at 6 ASoC SNP loci for validation (Table S25). NPC cells stably expressing dCas9 were prepared as described above and were replated into each well of a 12-well plate. Cells were transduced by 500 µl Crop-seq lenti-gRNA virus (with a cocktail pool of equal amounts of different gRNAs for the selected *cis* genes, or negative control gRNAs targeting EGFP). We included 4 biological replicates for each condition. The antibiotic selection of cells infected by lenti-gRNAs was performed as described above for scRNA-seq/CRISPRi. The cells were harvested and total RNAs were isolated (QIAGEN# 74134) for each condition, followed by qPCR

assay as described above (the qPCR assay IDs for the selected *cis*-genes were listed in Table. S24). Differential expression of each of the selected CROP-seq cis-target genes between cells infected by the ASoC SNP-specific gRNA and cells infected by negative control gRNAs was independently analyzed by unpaired Student's *t*-test with Welch's correction (n = 4 biological replicates).

8.1. CRISPR/Cas9 editing of ASoC SNP rs2027349 at *VPS45* locus

We carried out precise single nucleotide editing to examine the regulatory effect of ASoC SNP rs2027349 on *VPS45* in NPCs. Two different iPSC lines (from different donors) heterozygous for rs2027349 were used for editing to isogenic lines homozygous for risk allele or non-risk alleles. Each genotype had 2-3 iPSC lines. gRNAs for CRISPR/Cas9 editing were designed using an online tool (zlab.bio/guide-design-resources) and cloned into pSpCas9(BB)-2A-puro (PX459) v2.0 plasmid (Addgene #62988) following the protocol as described (*63*). Briefly, 4.5×10^5 iPSCs were replated onto 60 mm dishes as single cells 24 hrs before transfection. The next day, 3.0 µg of Cas9 plasmid carrying gRNA and 3 µg single-stranded donor oligonucleotide (ssODN) were co-transfected into cells using FuGENE HD (Promega) following vendor's instructions. 24 hrs post-transfection, cells were treated with 0.5 µg/ml puromycin for selection and media was refreshed with 0.25 µg/ml puromycin 48 hrs post-transfection (puromycin concentration were optimized for different lines). 7-10 days post-transfection, iPSC colonies were picked and further verified by Sanger sequencing at the ssODN-targeted sites. Colonies carrying the correct mutation were sub-cloned for purification. For sub-cloning, iPSCs were dissociated into single cells using Accutase (Thermofisher), and 5,000-8,000 cells were replated onto each 60 mm dish in mTeSR media with 5 µm ROCK inhibitor. 10-14 d post replating, colonies were picked, and the genotype of each colony was identified using Sanger sequencing. Only clones with the desired genotype were expanded for subsequent uses. The isogenic CRISPR-edited clones were all confirmed to be free of major off-target editing by Sanger sequencing of the top 5 predicted (zlab.bio/guide-design-resources) off-target sites. The isogenic iPSC lines were further differentiated into NPCs, and the local chromatin accessibility of OCR flanking the ASoC SNP rs2027349 was assayed in NPCs by ATAC-seq as described above and compared between the CRISPR-edited isogenic lines. The expression of *VPS45* (total RNA level and also for the major transcript isoform ENST00000369130) in isogenic CRISPR-edited iPSC-derived NPCs was measured by qPCR, and the differential expression between different genotypes (AA, AG, and GG) of the ASoC SNP rs2027349 was compared (see the statistical analysis below).

8.2. CRISPR/cas9 editing of ASoC SNP rs12895055 at *BCL11B* locus

We carried out precise single nucleotide editing to examine the regulatory effect of ASoC SNP rs12895055 on *BCL11B* in iN-Glut cells. CRISPR editing of rs12895055 was done as described above for *VPS45* locus. Two different iPSC lines (from two different donors) heterozygous for rs12895055 were used for editing to isogenic lines homozygous for the reference allele or the alternative allele. The decision of editing rs12895055 rather than its adjacent ASoC SNP rs11624408 in strong LD (r^2=0.98 in European population) was based on the following: (1) rs12895055 showed stronger ASoC than rs11624408, and (2) rs12895055 is within TF binding footprints of multiple TFs (EGR3, TCF4/TCF3 and SP4) of potential importance to neuropsychiatric disorders, and was predicted to have a stronger motif-disrupting effect than rs11624408 (in CENPB binding motif but with minimal effect) (Fig. S26C-D). The isogenic iPSC lines carrying different genotypes of rs12895055 were differentiated into iN-Glut cells as described above. The expression of *BCL11B* in iPSC-derived iN-Glut cells was measured by qPCR (we noted the expression of *BCL11B* in NPCs was too low to be able to reliably detected), and the differential expression between different genotypes of the ASoC SNP rs12895055 was compared (see the statistical analysis below).

8.3. Statistical analyses for CRISPR/Cas9 editing of ASoC SNPs

For analyzing the differential expression of the major transcript isoform (ENST00000369130) of *VPS45* between isogenic CRISPR-edited lines carrying different genotypes of ASoC SNP rs2027349, because we did not observe obvious line-to-line variation, we grouped all the different clones and independent experiments (cell cultures) of the two iPSC lines for each genotype and applied Student's *t*-test (with Welch's correction) (Fig. 4E). For analyzing the differential expression of *VPS45* (all transcript isoforms) and *BCL11B* in the CRISP/Cas9 SNP editing experiments described above, due to the obvious line-to-line expression variation, we first used linear regression to test the genotypic trend effect for each line separately (2 lines for each SNP), considering different clones and independent experiments (cell cultures) as biological replicates for each genotype. After confirming the absence of evidence of heterogeneity (comparing the β from the linear regression) between the two lines, we applied Fisher's method to combine *p*-values from the two lines to test for the significance of the ASoC SNP genotypic effects on *VPS45* and *BCL11B* expression, respectively.

Supplementary Figures

Question	Approach	Results	Figures
A. Are OCRs from iPSC-derived neurons (iNs) more similar to those in adult brains or fetal brains?	PCA of OCR peak signals across cell/tissue types; comparison of neuronal OCRs with adult brain data	OCRs of iNs are more similar to those at early neurodevelopmental stage; more OCRs in iNs than in adult brains	Figs. 1B, S7C, Table S3
B. Do genetic varaints affect chromatin accessibility in neurodevelopment (i.e., exhibiting ASoC) and in a cell-type-specific manner?	ASoC mapping in different cell types from the same set of indivudals; Storey's Pi1 analysis; comparing OCR and SNP effect sizes (allelic ratios) between cell types	ASoCs are prevalent and cell-type-specific; the specificty results from both cell-type specific OCRs and SNP effect size differences.	Figs. 1C-E, S9B, S10, S11
C. Do neuronal ASoC SNPs tend to be functional and regulate gene expression?	Intersecting ASoC with externa data, including epigenomic marks of enhancers, reporter gene assay, brain QTLs and Hi-C	ASoCs are enriched in brain enhancers and QTLs, functional in reporter gene assay; 59% of ASoCs have a cis-target annoated by eQTL, Hi-C contacts, and/or promoter	Figs. 2A-D, S12-14
D. Is alteration of TF binding a main mechanism of ASoC?	Correlation of motif changes of ASoC SNPs with their ASoC effects; Enrichment analysis of TF binding sites in ASoC	Genetic variants that alter TF-binding result in ASoC; Neuronal ASoC SNPs are enriched in TF-binding footprints important for neurodevelopment	Figs. 2E-G, S15-19
E. Are ASoC SNPs enriched with GWAS variants of neuropsychiatric diseases?	Test enrichment of GWAS risk variants in ASoC using a Bayesian statistical model (Torus)	ASoC shows strong enrichment (50-90 fold), compared with modest enrichment (<10) in OCRs and other annotations, e.g. conservation	Figs. 3A-C, S20
F. Which are the cis-target genes (thus likely the risk genes) of SZ-associated ASoC SNPs and how does the SNP risk allele affect a cis-target gene expression?	CRISPRi/single cell RNA-seq (CROP-seq) to systematically screen for cis-targets; combined with genetic fine-mapping to prioritize ASoC SNP for CRISPR/Cas9 editng; Examining the SNP allelic effect on cis-target gene in CRISPR-edited isogenic neural cells	CROP-seq nominated cis-target genes (some likely from distal regulation) for 10 ASoC sequences, most of which are validated by qPCR. CRISPR-editing of two prioritized ASoC SNPs demonstrated the allelic effects on VPS45 and BCL11B, suggesting them as likely SZ risk genes	Figs. 4A-F, S21-26

Fig. S1: Overview of the study questions, approaches/experiments and main results

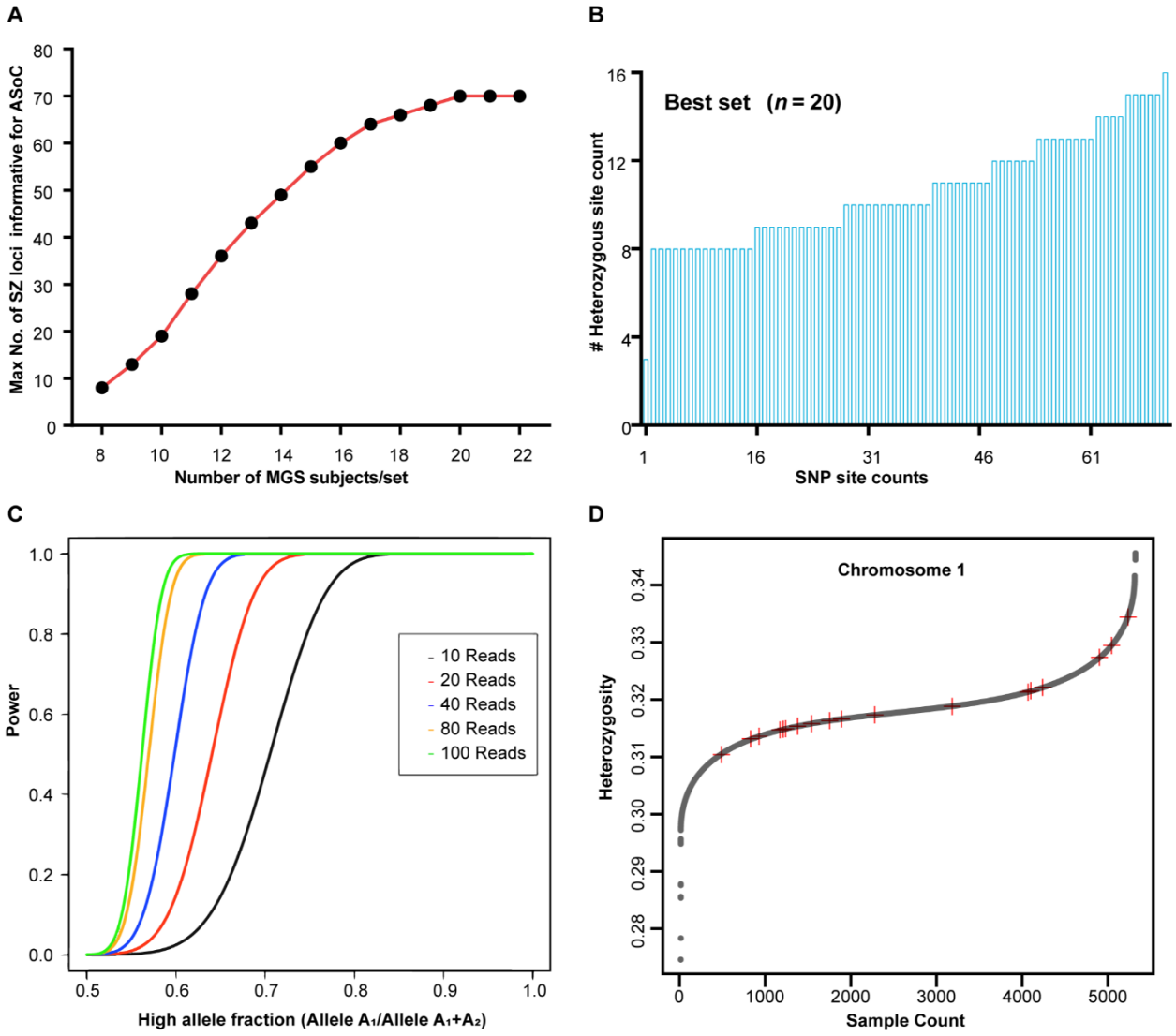

Fig. S2: Selection and evaluation of the 20 MGS subjects for deriving iPSC lines.

(A) The maximum number of heterozygous (*i.e.*, informative for allele-specific assay) SZ loci as a function of the number of the selected MGS subjects. Each of the 108 SZ loci was indexed by a genome-wide significant SNP with $r^2 \geq 0.8$ with other proxy SNPs (a total of 3,507 index and proxy SZ GWAS SNPs from Table S4 of (*5*)), and we used only the 70 SZ GWAS SNPs that were genotyped in the MGS GWAS study (*9*) for the sample selection. The curve reached its saturation point at approximately 20 subjects. **(B)** Count of heterozygous subjects at each of the 70 SZ GWAS SNP site in the selected 20 individuals (best set of 20 MGS subjects). **(C)** The power to detect ASoC variants (at ~70/108 SZ GWAS risk loci) as a function of the detected high (more frequent) allelic ratio for the 20 selected MGS subjects. The power calculation was based on the binomial test to detect ASoC SNP. **(D)** The distribution of the heterozygosity of the 20 selected subjects (red cross) together with all other MGS subjects (only chr1 SNPs were shown).

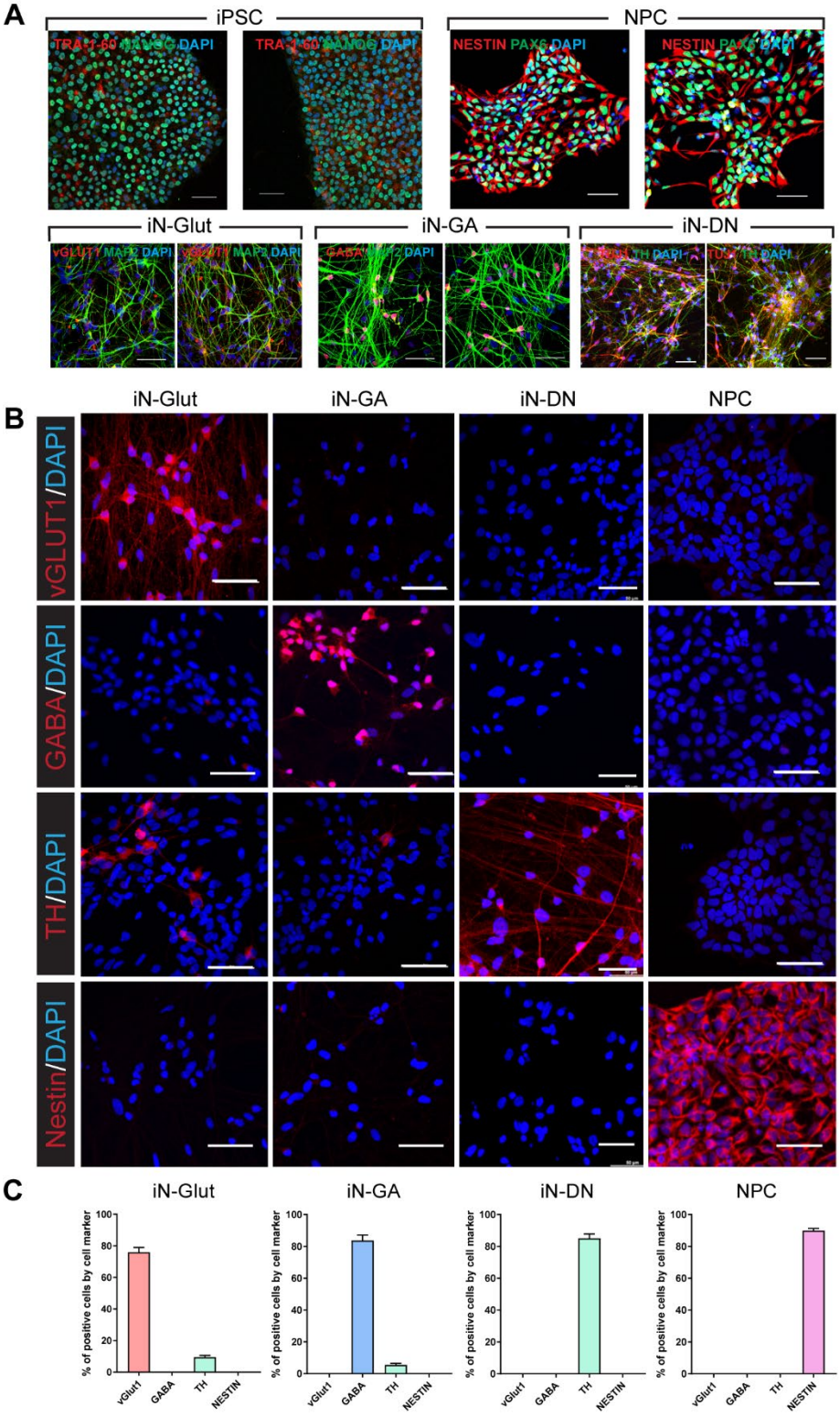

Fig. S3: Morphological characterization of iPSC lines and iPSC-differentiated neuronal cells.

(A) Representative immunofluorescence (IF) staining images of iPSC and neuronal cells stained with antibodies against cell-type-specific markers. Scale bar: 50 µm. **(B)** Representative images

used to estimate the purity of each neural cell type (iN-Glut, iN-GA, iN-DN and NPC). Cells were stained for both cell-type-specific marker and markers specific for other cell types. Scale bar: 50 μm. **(C)** Purity of the cells used for ATAC-seq. Shown are the percentages of neuronal cells that are stained positive with their corresponding type-specific markers (approximately 100 cells per field of view and 5-10 fields were used for each cell line). Percentages of other "contaminating" types of cells were also shown for each cell type. Cell-type-specific markers are: iPSC: TRA-1-60$^+$/NANOG$^+$; NPC: NESTIN$^+$/PAX6$^+$; iN-Glut: vGLUT1$^+$/MAP2$^+$; iN-GA: GABA$^+$/MAP2$^+$; iN-DN: TUJ1$^+$/TH$^+$.

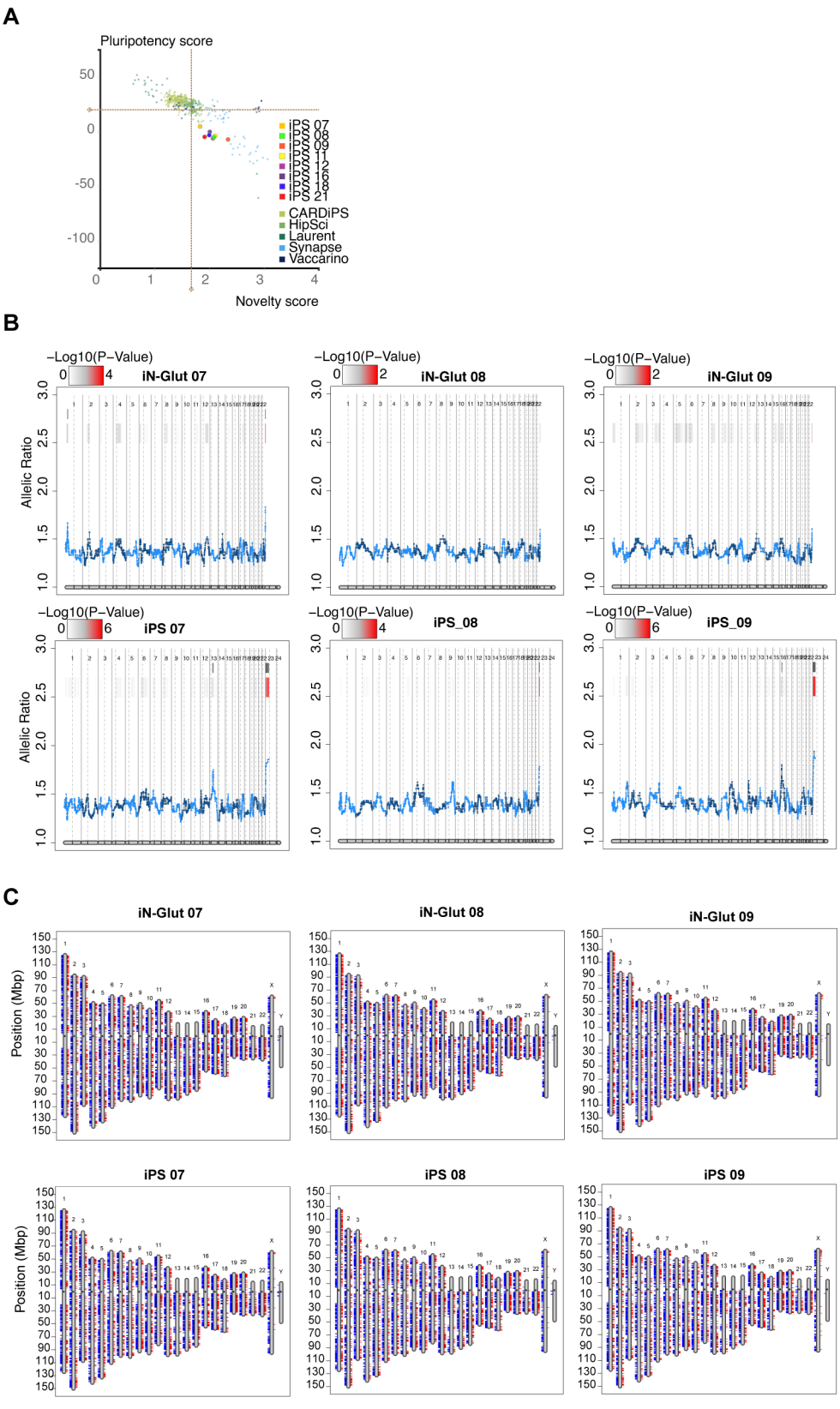
A
Pluripotency score
Novelty score
iPS 07
iPS 08
iPS 09
iPS 11
iPS 12
iPS 16
iPS 18
iPS 21
CARDiPS
HipSci
Laurent
Synapse
Vaccarino
B
−Log10(P−Value)
iN-Glut 07
iN-Glut 08
iN-Glut 09
iPS 07
iPS_08
iPS_09
Allelic Ratio
C
Position (Mbp)

Fig. S4: Molecular and genetic characterization of iPSC lines and iNs using RNA-seq data.

(A) Pluripotency/Novelty score of iPSC of the core-8 iPSC cell lines by using Pluritest (with RNA-seq data as inputs) (*28*). The core-8 iPSC lines were clustered and compared to iPSC lines from other cohorts (provided by Pluritest). **(B)** eSNP-Karyotyping (*29*) using RNA-seq data of iPSC lines and their corresponding neuronal cell types to confirm the absence of chromosome abnormality. Shown are moving average plots of RNA-seq intensities of called SNPs along the chromosomes (iPSC cells and iN-Glut neurons for 3 cell lines are shown as examples). No chromosome aneuploidy in iPSC lines or iN-Glut neurons were identified. Color bars represent the scale of $-\log_{10}$ value (FDR-calibrated *P* value) of allelic ratio differences of a SNP marker. **(C)** Representative per-chromosome recombination map of the corresponding cell lines and neural cultures in our study (from eSNP-Karyotyping).

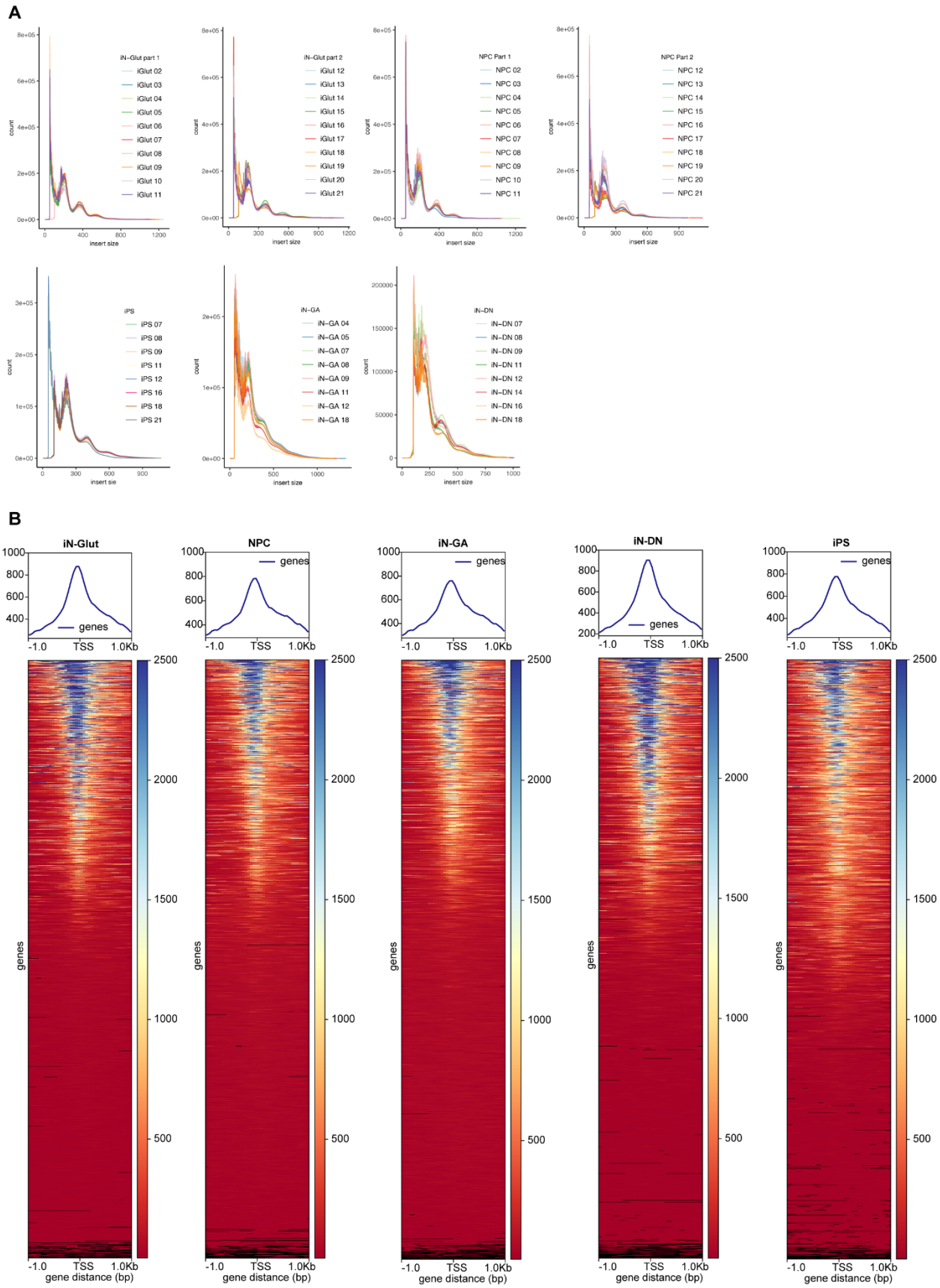
A
iN−Glut part 1
iGlut 02
iGlut 03
iGlut 04
iGlut 05
iGlut 06
iGlut 07
iGlut 08
iGlut 09
iGlut 10
iGlut 11
iN−Glut part 2
iGlut 12
iGlut 13
iGlut 14
iGlut 15
iGlut 16
iGlut 17
iGlut 18
iGlut 19
iGlut 20
iGlut 21
NPC Part 1
NPC 02
NPC 03
NPC 04
NPC 05
NPC 06
NPC 07
NPC 08
NPC 09
NPC 10
NPC 11
NPC Part 2
NPC 12
NPC 13
NPC 14
NPC 15
NPC 16
NPC 17
NPC 18
NPC 19
NPC 20
NPC 21
iPS
iPS 07
iPS 08
iPS 09
iPS 11
iPS 12
iPS 16
iPS 18
iPS 21
iN−GA
iN−GA 04
iN−GA 05
iN−GA 07
iN−GA 08
iN−GA 09
iN−GA 11
iN−GA 12
iN−GA 18
iN−DN
iN−DN 07
iN−DN 08
iN−DN 09
iN−DN 11
iN−DN 12
iN−DN 14
iN−DN 16
iN−DN 18
count
insert size
B
iN-Glut
NPC
iN-GA
iN-DN
iPS
genes
-1.0
TSS
1.0Kb
gene distance (bp)

Fig. S5: QC metrics of ATAC-seq data.

(A) Histogram of the insert sizes of all ATAC-seq samples, showing the expected distribution pattern of transposase insertion at specific sizes (in the unit length of nucleosomes; X-axis). iN-Glut and NPC samples (n=20 per cell type) were plotted with two graphs for better resolution. **(B)** Chromatin accessibility in regions proximal to the TSS of core promoters (± 1kb) in different cell types, showing the library size-normalized overall ATAC-seq read intensity (top) and per-gene heatmap (bottom). The strong signal to noise ratio of the ATAC-seq peak intensity around the TSS site indicated a good quality of ATAC-seq data. For each cell type, reads from core-8 samples were pooled together to generate the density map. Color bar indicates the density of reads at each position proximal to the TSS site. Black regions represent a loss of signal at the corresponding coordination.

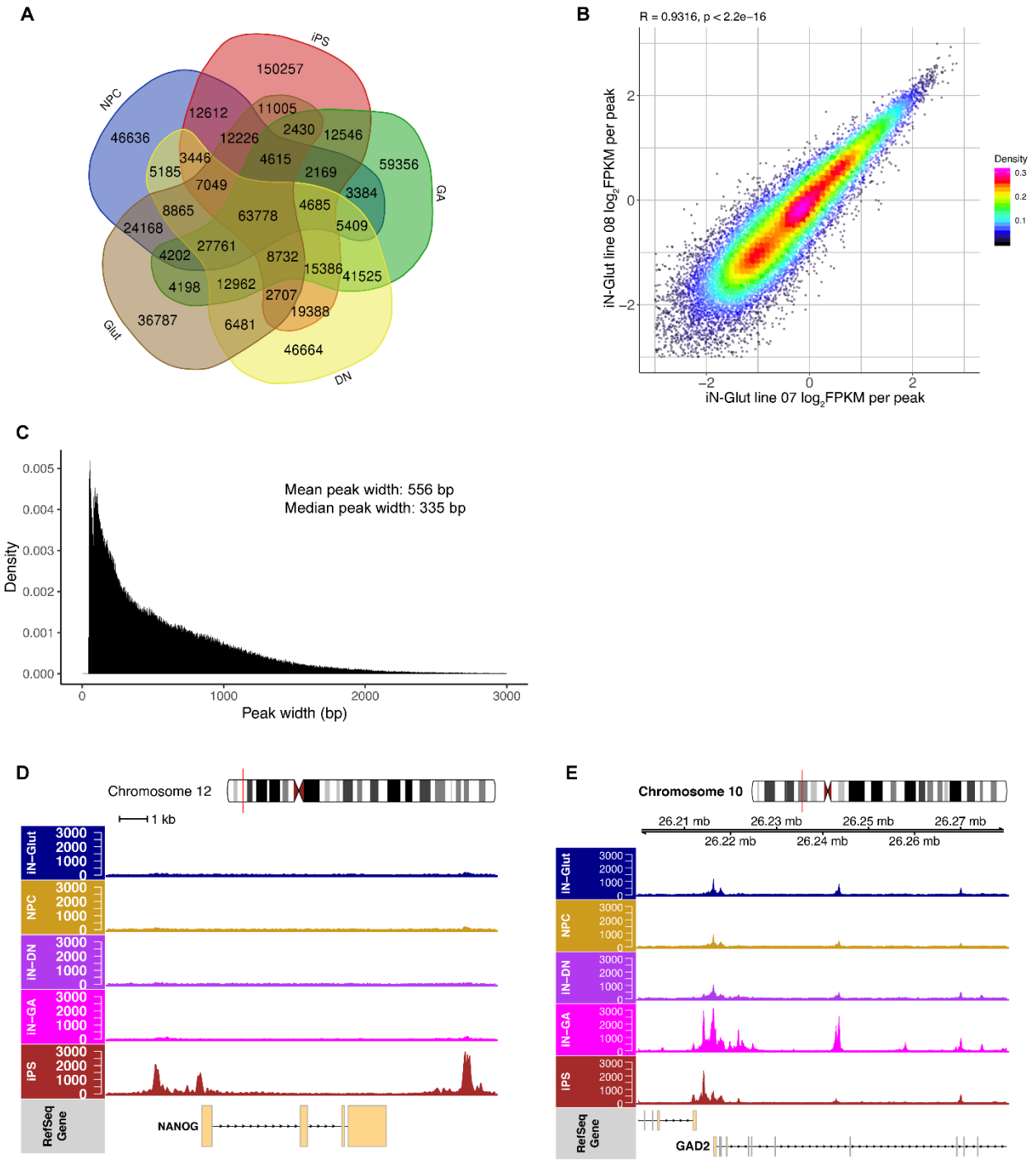

Fig. S6: ATAC-seq open chromatin region (OCR) peak calling and QC.

(A) Venn diagram of the number of OCRs (peaks) called in different cell types. MACS2 was used for peak calling (FDR < 0.05) with the pooled reads of the core-8 lines for each cell type. **(B)** Pearson correlation of the ATAC-seq peak reads (in called OCRs) between cell lines within the same cell type. Shown are normalized read counts of the same peak sets in two different lines (iN-Glut 07 and iN-Glut 08 as an example; considered as biological replicates). All pair-wise correlation heatmap is shown in Fig. S7A. **(C)** Histogram of ATAC-seq peak width distribution of

all tested samples. X-axis: peak width in bp. **(D)** ATAC-seq peaks of *NANOG* locus in different cell types, showing the expected iPSC-specific OCR at *NANOG* locus. **(E)** ATAC-seq peaks of *GAD2* locus in different cell types, showing the expected stronger chromatin accessibility in iPSC-derived GABAergic neurons (iN-GA).

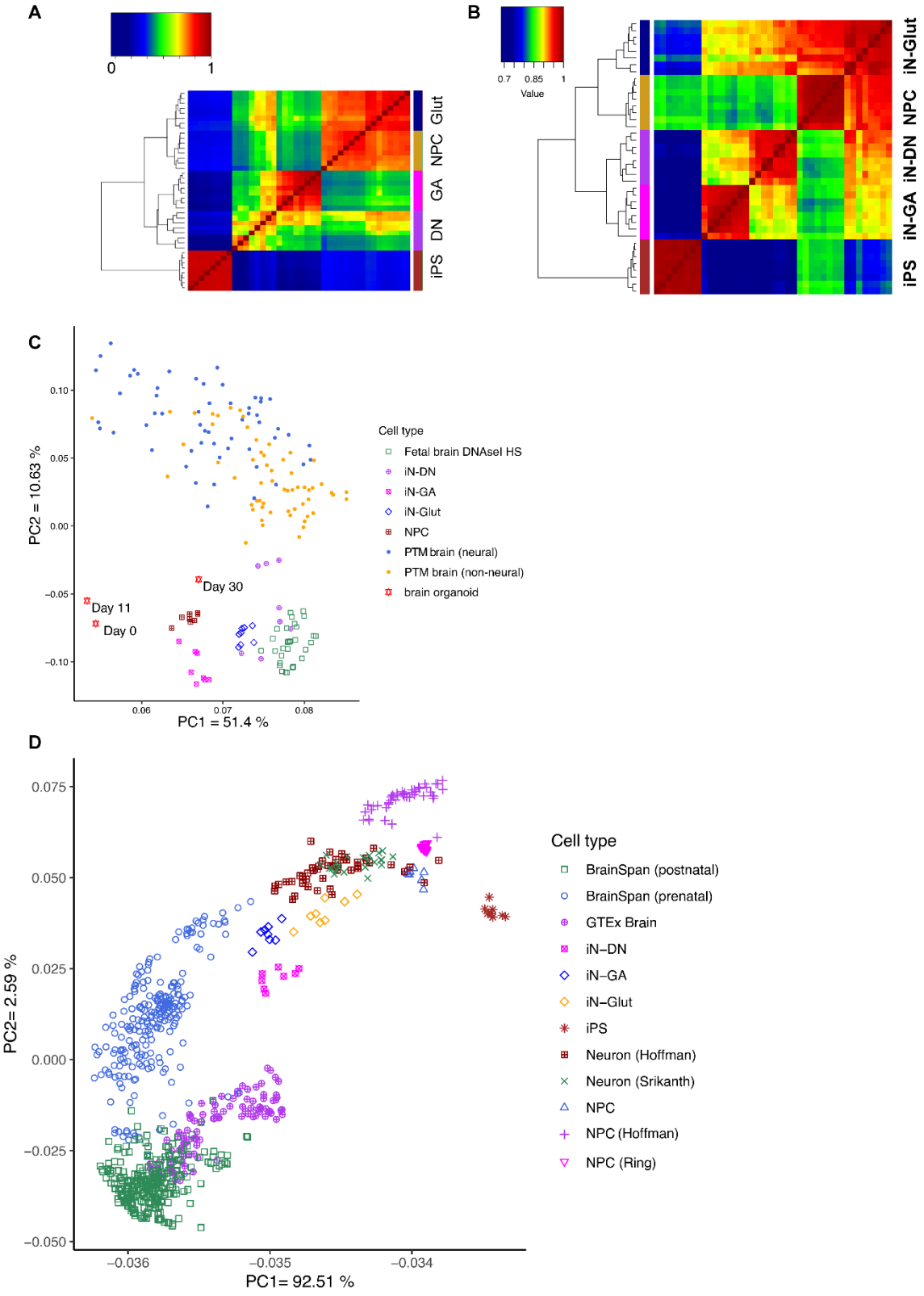
A
0
1
Glut
NPC
GA
DN
iPS
B
0.7
0.85
1
Value
iN-Glut
iN-NPC
iN-DN
iN-GA
iPS
C
PC2 = 10.63 %
PC1 = 51.4 %
Day 30
Day 11
Day 0
Cell type
Fetal brain DNAsel HS
iN-DN
iN-GA
iN-Glut
NPC
PTM brain (neural)
PTM brain (non-neural)
brain organoid
D
PC2= 2.59 %
PC1= 92.51 %
Cell type
BrainSpan (postnatal)
BrainSpan (prenatal)
GTEx Brain
iN–DN
iN–GA
iN–Glut
iPS
Neuron (Hoffman)
Neuron (Srikanth)
NPC
NPC (Hoffman)
NPC (Ring)

Fig. S7: Hierarchical clustering and PCA of different cell types by using ATAC-seq peak reads and RNA-seq gene expression levels.

(A) Hierarchical clustering and heatmap of normalized ATAC-seq reads within OCR peaks of different cell types. OCR peaks were called from the pooled reads of from core-8 samples of each cell type. **(B)** Hierarchical clustering and heatmap of RNA-seq gene expression levels of different cell types. Gene expression levels were averaged from core-8 samples of each cell type. **(C)** PCA analysis of ATAC-seq samples of NPCs and iNs (i.e., without iPSCs) in comparison with multiple publicly available ATAC-seq or DNaseI-seq datasets on iPSC-derived neural cells or brain tissues. **(D)** PCA analysis of RNA-seq samples of the core-8 samples in comparison with multiple publicly available RNA-seq datasets on iPSC-derived neural cells or brain tissues. A consensus list of 15,751 genes was used in both hierarchical clustering in (B) and PCA analysis in (D).

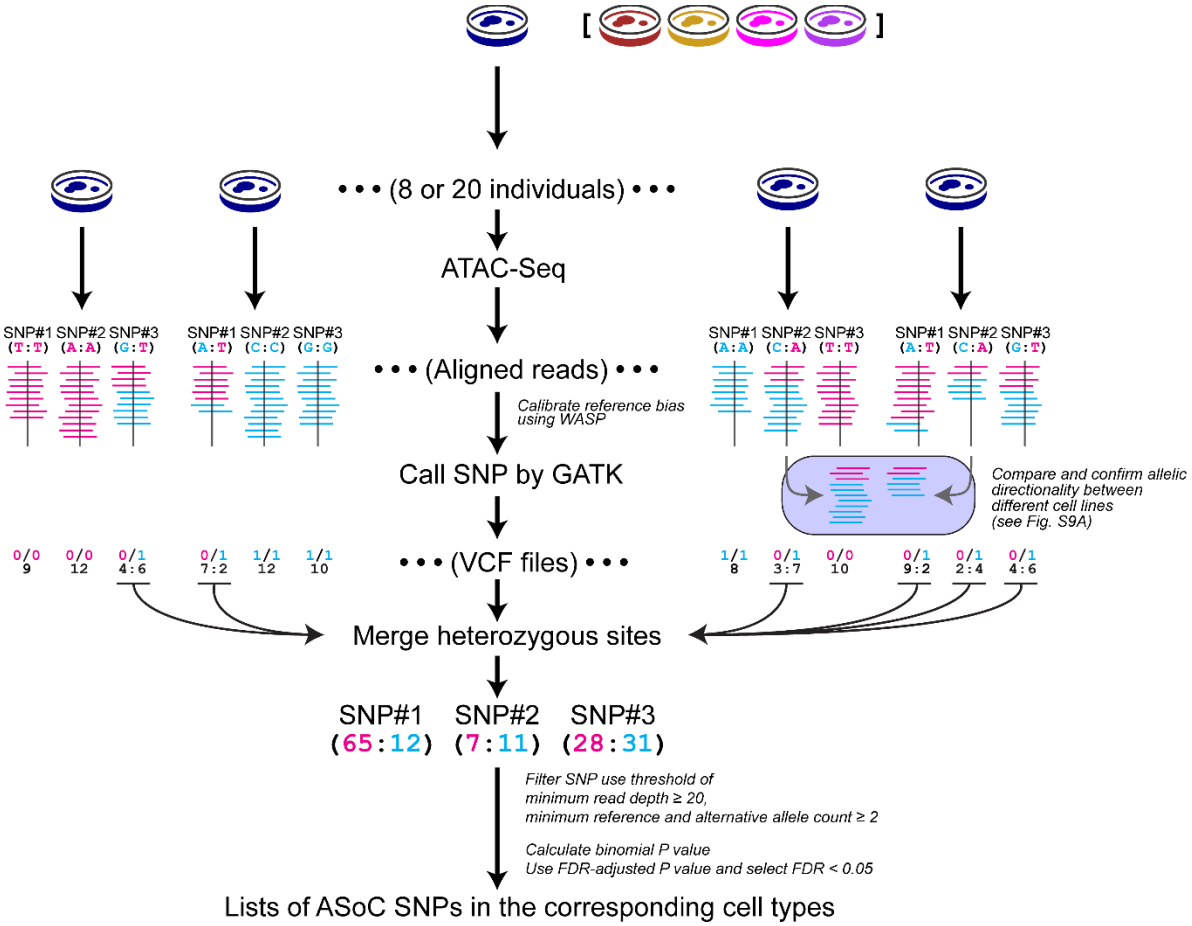

Fig. S8: Flowchart of ASoC SNP calling.

For each cell type, ATAC-seq was carried out for each cell line (a total of 8 individuals for iN-GA, iN-DN, and iPS cells, and 20 individuals for iN-Glut and NPC for increased statistical power). ATAC-seq reads were aligned to hg38 genome and further filtered by WASP to eliminate allelic bias towards reference alleles. GATK (v4.0) was used to call SNPs for each sample individually, and only heterozygous sites were retained in the output VCF files. The generated VCF files were subsequently merged by cell type for heterozygous SNPs, and a read-depth cut-off threshold (n<20) was applied to eliminate SNPs within weak OCR peaks. A binomial test was applied to test allelic imbalance of the ATAC-seq reads flanking a heterozygous SNP. Heterozygous SNPs with FDR < 0.05 in the binomial test of allelic imbalance of ATAC-seq reads were considered as ASoC SNPs.

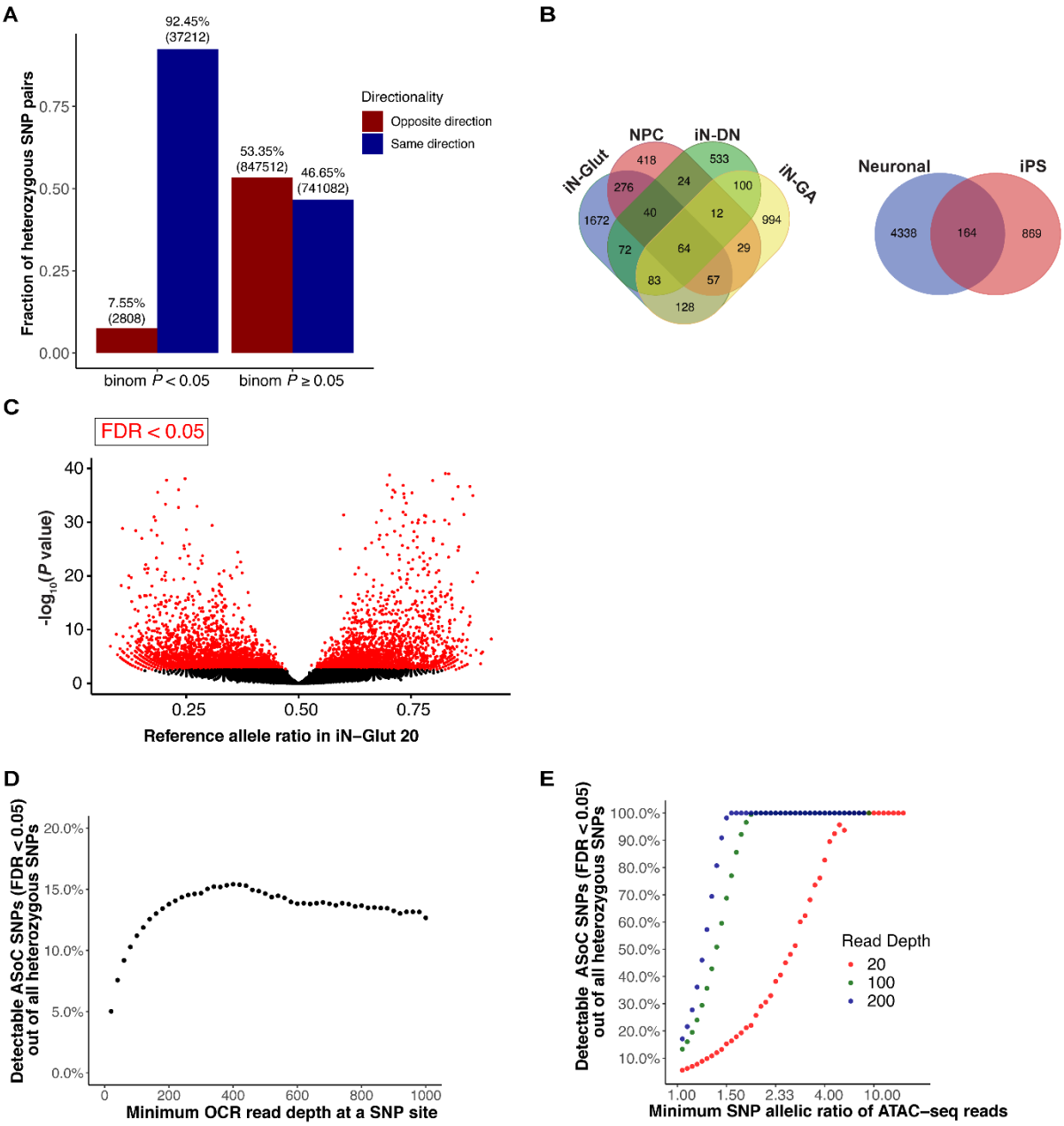
A
92.45%
(37212)
53.35%
(847512)
46.65%
(741082)
7.55%
(2808)
Fraction of heterozygous SNP pairs
0.00
0.25
0.50
0.75
binom P < 0.05
binom P ≥ 0.05
Directionality
Opposite direction
Same direction
B
NPC
iN-DN
iN-Glut
iN-GA
418
533
276
24
100
1672
40
12
994
72
64
29
83
57
128
Neuronal
iPS
4338
164
869
C
FDR < 0.05
-log10(P value)
0
10
20
30
40
0.25
0.50
0.75
Reference allele ratio in iN−Glut 20
D
Detectable ASoC SNPs (FDR < 0.05) out of all heterozygous SNPs
0.0%
5.0%
10.0%
15.0%
20.0%
0
200
400
600
800
1000
Minimum OCR read depth at a SNP site
E
Detectable ASoC SNPs (FDR < 0.05) out of all heterozygous SNPs
10.0%
20.0%
30.0%
40.0%
50.0%
60.0%
70.0%
80.0%
90.0%
100.0%
Read Depth
20
100
200
1.00
1.50
2.33
4.00
10.00
Minimum SNP allelic ratio of ATAC−seq reads

Fig. S9: ASoC SNPs called in each cell type (by pooling) and ASoC detection sensitivty.

(A) Inter-individual concordance of the directionality of allelic imbalance at heterozygous SNP sites. Shown are pairs of the iN-Glut-20 samples with ASoC in both samples (binomial $p < 0.05$), and the pairs with non-ASoC in both samples (binomial $p > 0.05$). **(B)** Venn diagram of ASoC SNP numbers ($n = 8$ samples per cell type). **(C)** Volcano plot showing the reference allelic ratio of all heterozygous SNPs (X-axis) and their corresponding *p*-values of ASoC testing (Y-axis; in -log10 scale) from 20 iN-Glut samples. Red: 5,611 ASoC SNPs with FDR<5%. **(D)** Summary plot shows the percentage of heterozygous SNPs that were called as ASoC SNPs for OCRs at different sequencing read-depths. iN-Glut-20 ASoC dataset was used. The total allele counts of a called heterozygous SNP was used as a proxy of OCR read-depth. **(E)** Summary plot shows the sensitivity of detecting ASoC SNPs with different effect sizes (allelic ratios) at a given OCR read-depth (>20, >100 or >200). iN-Glut-20 ASoC dataset was used.

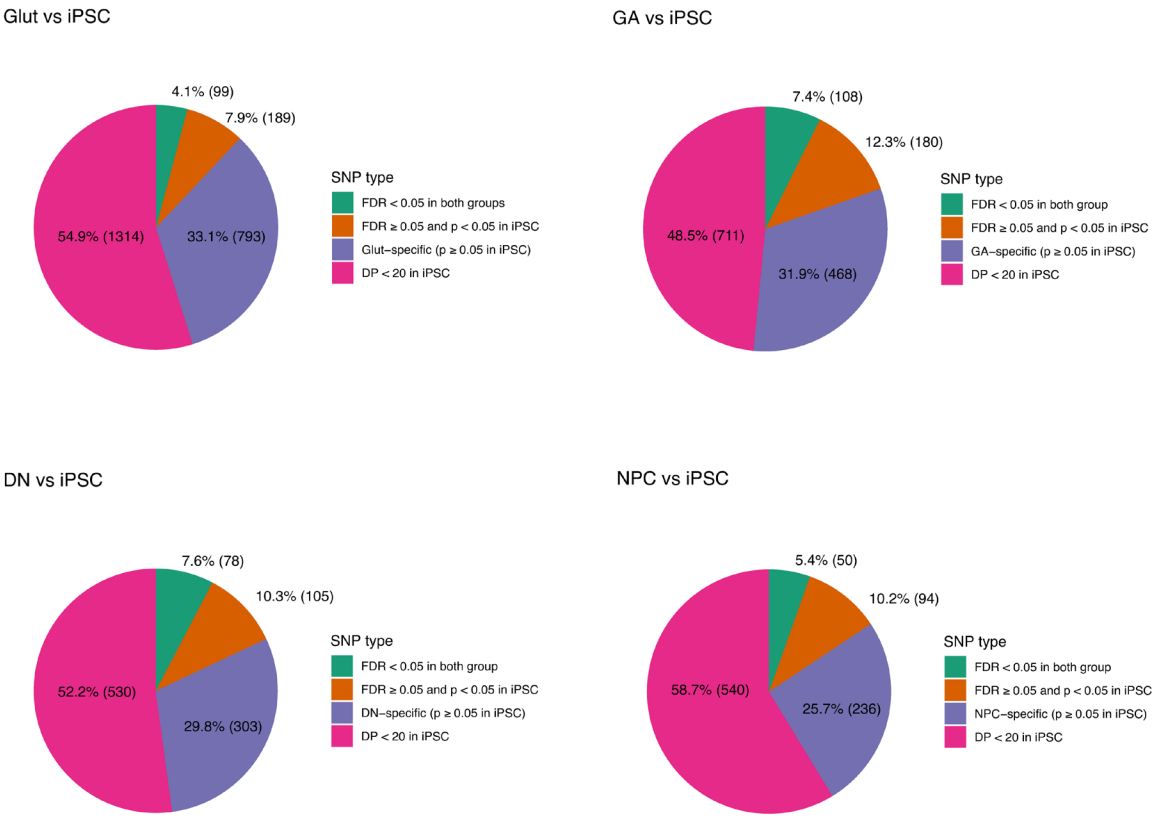

Fig. S10: Comparison of ASoC SNPs between neurons and iPSCs.

The analyses were done on core-8 samples of each cell type. For each of the four neuronal cell types, we ascertained on ASoC SNPs (FDR < 0.05), and then divided them into 4 types (colors) according to their ASoC FDR/p-values or read depth in iPSC. Pie charts show the proportions of 4 SNP types in each neuronal cell type compared to iPSC.

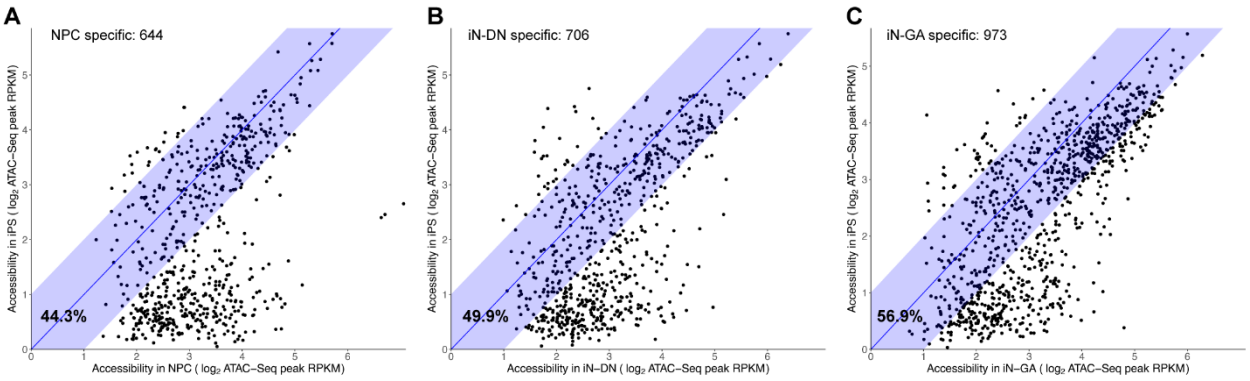

Fig. S11: Chromatin accessibility of OCR peaks (neuron vs. iPSC) at neuron-specific ASoC SNP sites.

The analyses were performed on data from the core-8 samples of each cell type. ATAC-seq peaks called from each of the five cell types were merged together to create a unified set of peaks, and for each of the 40 samples, the number of ATAC-seq reads that fell into each peak was counted and converted to RPKM after normalization by the total number of reads and peak length. For each neuronal cell type, neuron-specific ASoC SNPs were defined as those with an FDR < 0.05 (binomial test) in the neuronal cell type while having either a read depth (DP) < 20 or a nominal *p*-value > 0.05 (binomial test) in iPSC. For all the OCR peaks that contain neuron-specific ASoC SNPs, we plotted the mean of their $\log_2$(RPKM) counts across core-8 neuronal samples versus the mean $\log_2$(RPKM) across core-8 iPSC samples. The blue shaded area along the diagonal contains OCR peak where ATAC-seq read difference between iPSC and neuronal cell type is within two-fold (*i.e.*, comparable OCR signal in our definition). **(A)** OCR peaks for NPC specific ASoC SNPs. **(B)** OCR peaks for iN-DN specific ASoC SNPs. **(C)** OCR peaks for iN-GA specific ASoC SNPs.

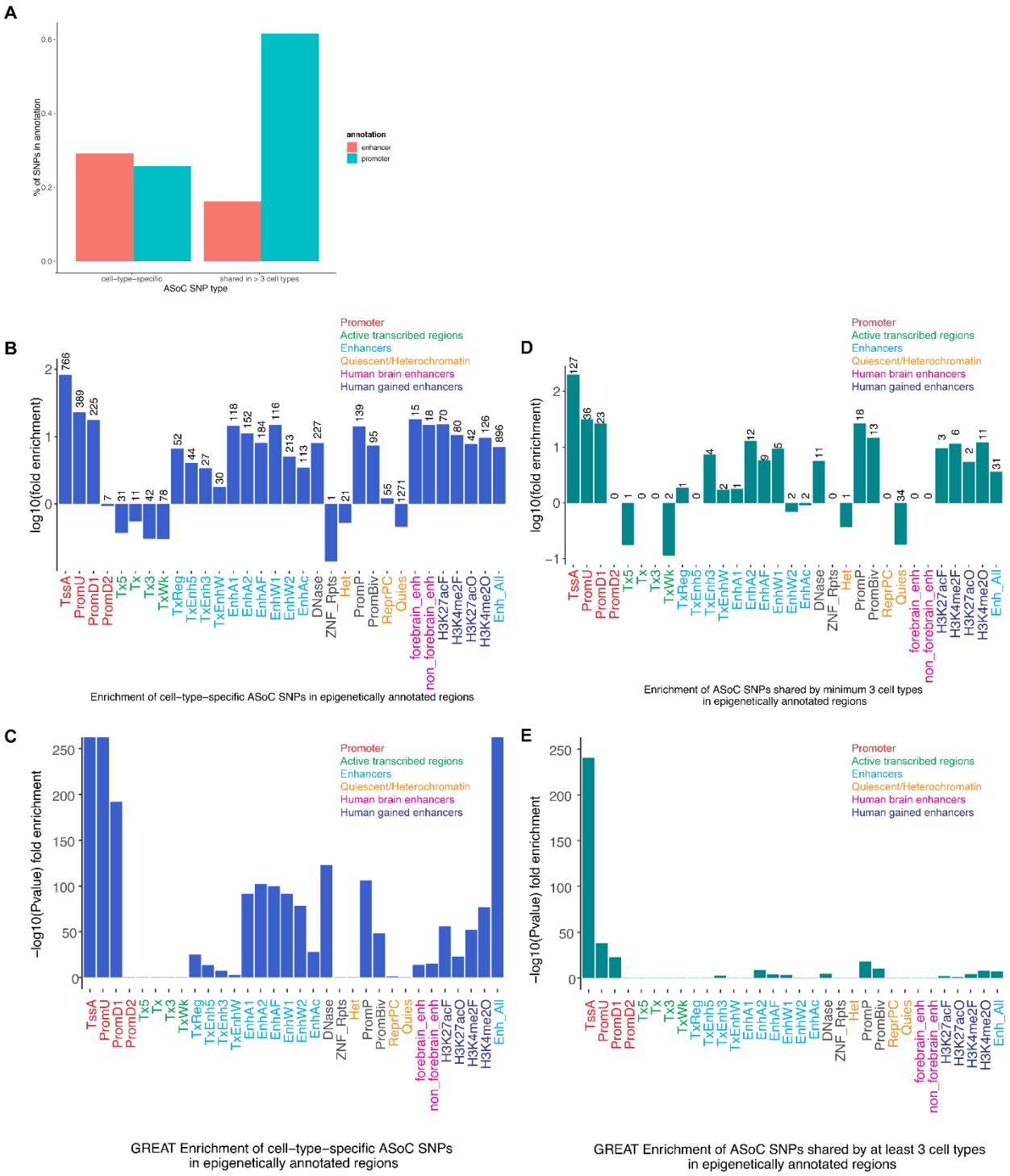

Fig. S12: Enrichment of cell-type-specifc vs shared ASoC SNPs in epigenomic features.

(A) Bar graph showing the percentages of cell-type-specific and shared ASoC SNPs within promoters (TssA, PromU, PromD1 and PromD2) and enhancers (TxReg, TxEnh5, TxEnh3, TxEnhW, EnhA1, EnhA2, EnhAF, EnhW1, EnhW2, EnhAc, DNase) annotated by chromatin-state-based epigenomic features. **(B)** Fold enrichment of cell-type-specific ASoC SNPs in different types of chromatin-state-based epigenetically annotated promoters, enhancers, and other

functional regions. The number above each bar is the number of ASoC SNPs in each epigenomic feature. A combined enhancer category "Enh_All" = EnhA1+ EnhA2 + EnhAF + EnhW1 + EnhW2 + EnhAc. **(C)** Enrichment -$\log_{10}$ *P* value of cell-type-specific SNPs in (B). **(D)** Fold of enrichment of ASoC SNPs that are shared in ≥ 3 cell types in epigenetically annotated genomic features. **(E)** Enrichment -$\log_{10}$ *P* value of cell-type-shared SNPs in (D). Enrichment test background was genome size for (B) to (E), and the enrichment *P*-values in (C) and (E) were derived by binomial test implemented in GREAT (*4, 47*).

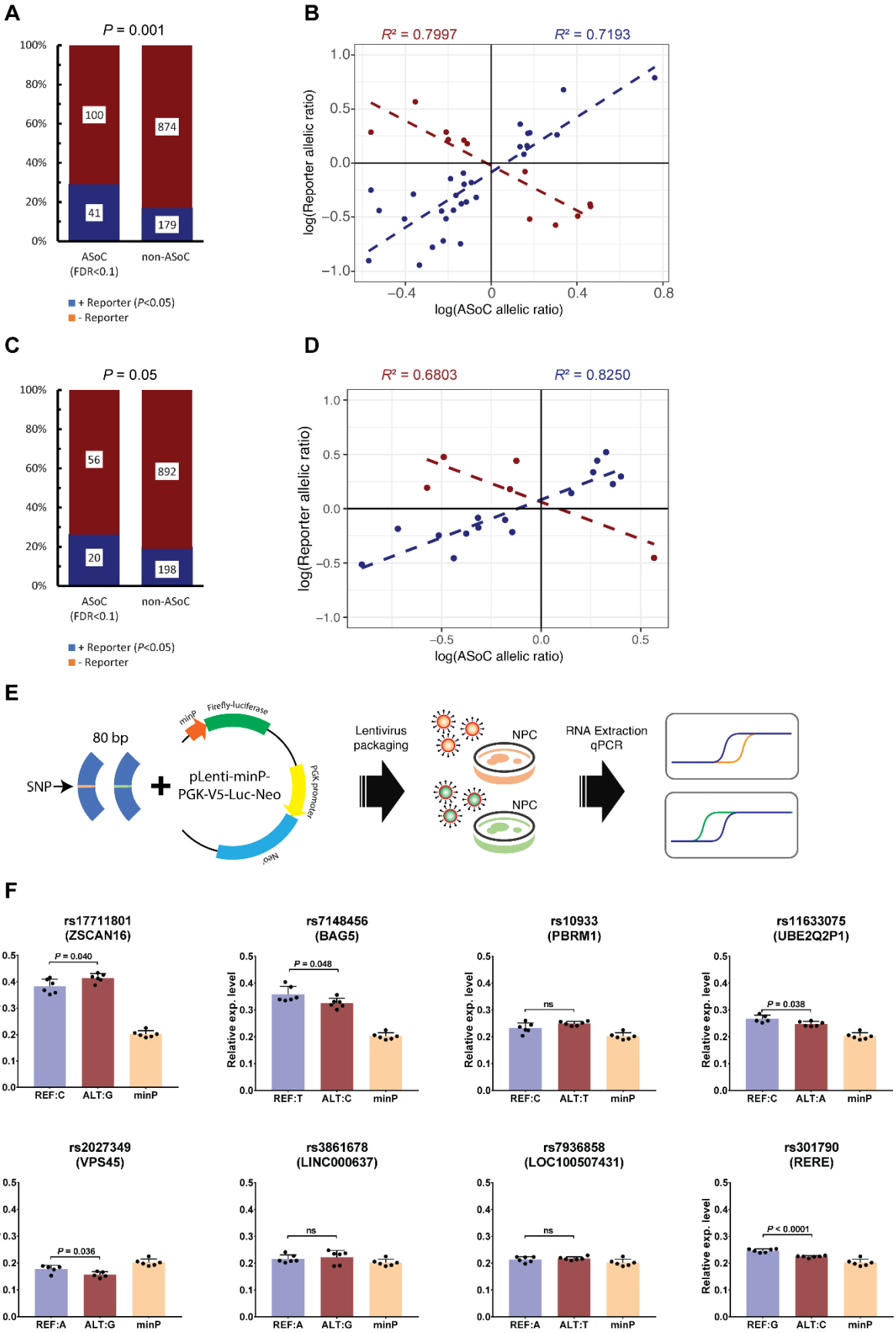
A
P = 0.001
100%
80%
60%
40%
20%
0%
100
874
41
179
ASoC (FDR<0.1)
non-ASoC
+ Reporter (P<0.05)
- Reporter
B
R² = 0.7997
R² = 0.7193
log(Reporter allelic ratio)
log(ASoC allelic ratio)
C
P = 0.05
56
892
20
198
D
R² = 0.6803
R² = 0.8250
E
80 bp
SNP
pLenti-minP-PGK-V5-Luc-Neo
Firefly-luciferase
minP
PGK promoter
Neo
Lentivirus packaging
NPC
RNA Extraction qPCR
F
rs17711801 (ZSCAN16)
P = 0.040
REF:C
ALT:G
minP
rs7148456 (BAG5)
P = 0.048
REF:T
ALT:C
rs10933 (PBRM1)
ns
REF:C
ALT:T
rs11633075 (UBE2Q2P1)
P = 0.038
REF:C
ALT:A
rs2027349 (VPS45)
P = 0.036
REF:A
ALT:G
rs3861678 (LINC000637)
ns
REF:A
ALT:G
rs7936858 (LOC100507431)
ns
REF:A
ALT:T
rs301790 (RERE)
P < 0.0001
REF:G
ALT:C
Relative exp. level

Fig. S13: Reporter gene analyses to validate the effects of ASoC SNPs on enhancer/promoter activity.

(A) Percentage of iN-Glut ASoC SNPs or non-ASoC SNPs that are found to alter reporter gene expression (blue portion of the bar) in a published massive parallel reporter assay (MPRA) SNP data set from lymphoblastoid cell lines (LCLs) (*15*). ASoC SNPs (FDR<0.1, binomial test) and non-ASoC SNPs in iN-Glut-20 were used to intersect with the list of SNPs (n=32,373) that were tested for SNP allelic effect on reporter gene expression in LCLs. Fisher's exact test was used to estimate the enrichment of ASoC SNPs (Fold_enrichment = 2, p = 0.001, vs. non-ASoC SNPs) that alter reporter gene expression (+reporter). **(B)** Correlation of the allelic effects of iN-Glut ASoC SNPs on chromatin accessibility and on reporter gene expression. Allelic ratios were expressed as the ATAC-seq reads (X-axis) or reporter gene expression (Y-axis) of ref vs. alternative allele. Each dot represents a SNP. Blue line: positive correlation; red line: negative correlation. **(C)** Percentage of NPC ASoC SNPs or non-ASoC SNPs that are found to alter reporter gene expression (blue portion of the bar) in a published massive parallel reporter assay (MPRA) SNP data set from lymphoblastoid cell lines (LCLs) (*15*). Analysis was done as in (A). **(D)** Correlation of the allelic effects of NPC ASoC SNPs on chromatin accessibility and on reporter gene expression. Analysis was done as in (B). **(E)** Schematic of the reporter gene assay used for evaluating the regulatory effect of ASoC SNPs in NPCs. DNA segments (±40 bp proximal to ASoC SNPs) for both alleles were individually cloned into a modified pLenti-PGK-V5-LUC-Neo construct. The vector encodes the reporter gene luciferase (LUC) and Neo (G418 resistance gene). The expression levels of LUC of both alleles (normalized by Neo expression) were compared to estimate the regulatory effect of a ASoC SNP on promoter/enhancer activity of the DNA sequence that flank the ASoC SNP. **(F)** Bar graph of qPCR results from the reporter gene assay shows the relative LUC expression for each allele of an assayed ASoC SNP. Reporter gene expression from a separate construct carried a minimal promoter (denoted as minP) shows the basal promoter activity level in NPCs. 8 ASoC SNPs (adjacent genes in parenthesis) associated with SZ were assayed. Each point represents an independent biological replicate. p-values testing the allelic effect on reporter expression are computed from unpaired Student's t-test with Welch's correction. ns: not significant. Error bars: standard deviation (SD) for 5-6 biological replicates.

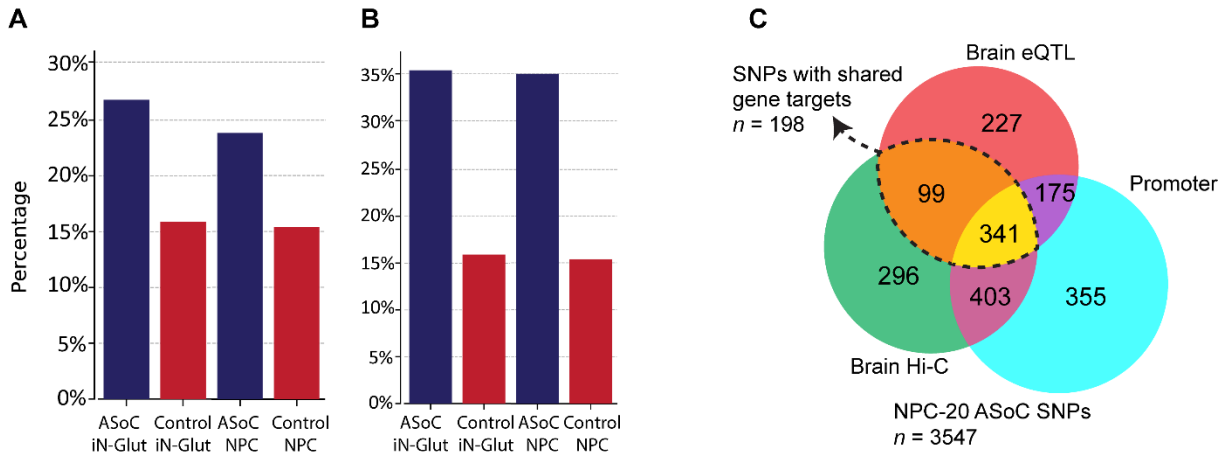

Fig. S14: Neuronal ASoC SNPs (iN-Glut-20 and NPC-20) that are also annotated as brain eQTLs, brain/neuronal Hi-C data, and promoters.

(A) Percentage of iN-Glut-20 and NPC-20 ASoC SNPs that are CommonMind brain eQTLs. The control SNP sets are random SNPs in the genome (red bar). **(B)** The same as in (A) except by restricting the ASoC or control SNPs to those within PsychENCODE adult brain OCRs (*3*). **(C)** Venn diagram showing the overlap of NPC-20 ASoC SNPs with brain eQTLs, brain/neuronal Hi-C *cis*-targets and promoters (Table S15). Area within the dashed line shows ASoC SNPs that are eQTLs and also mapped to Hi-C *cis*-targets.

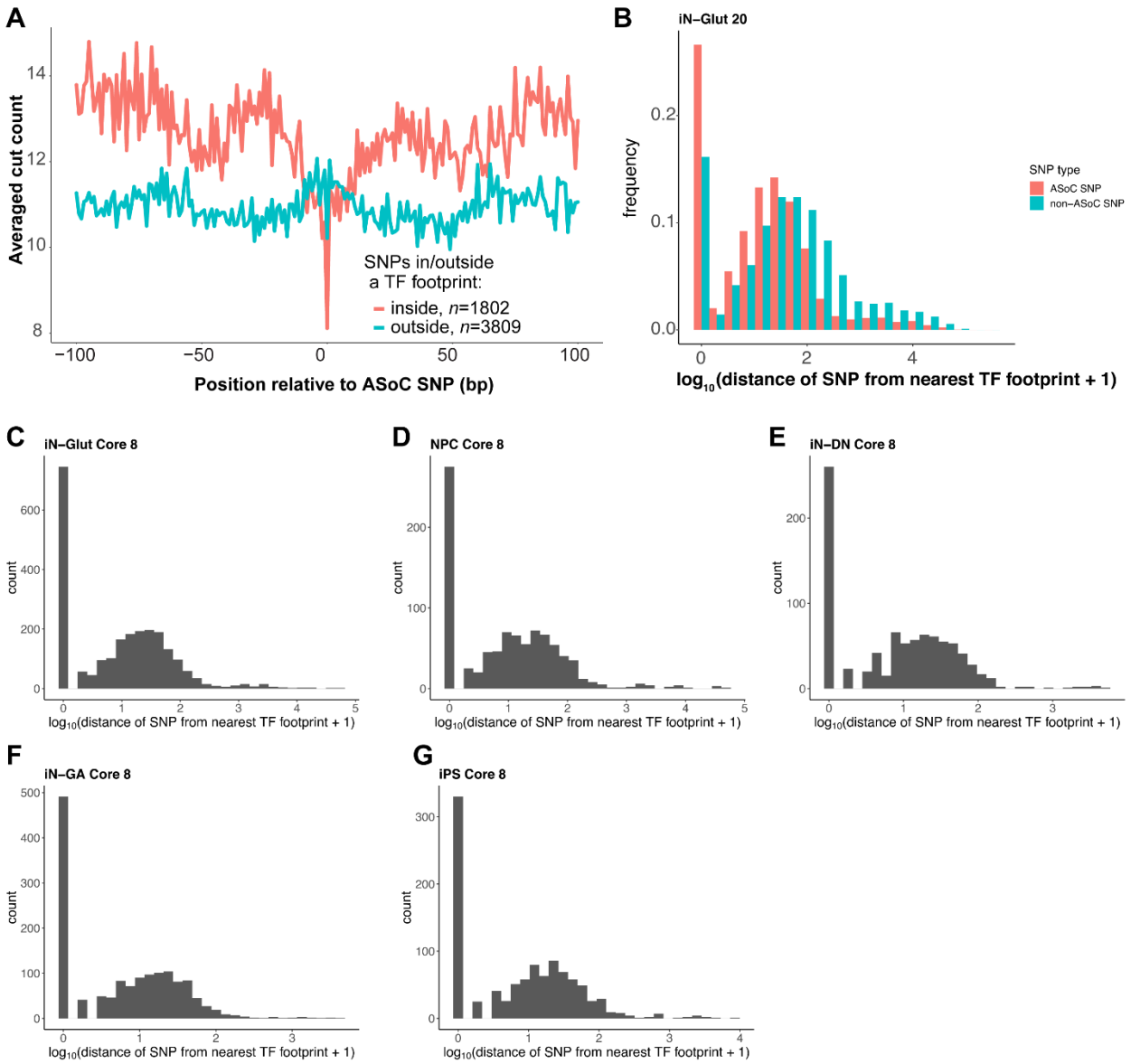

Fig. S15: TF footprint analysis using ATAC-seq data in each cell type.

(A) Averaged ATAC-seq cleavage profiles around ASoC SNPs inside (red, n = 1,802) or outside (blue, n = 3,809) predicted TF footprints. 23,858 out of 106,107 non-ASoC SNPs are inside predicted TF-footprints. The ATAC-seq cleavage profiles were generated by piling up the 5' ends of ATAC-seq reads in 200 bp windows centered around the ASoC SNPs. Note the strong dip of the average cut count at position 0 of a TF footprint. **(B)** Histogram of the iN-Glut-20 ASoC SNPs (vs non-ASoC SNPs) based on their distances (bp) relative to the nearest TF footprint. A total of 1,004,314 non-redundant TFBSs for a set of 522 TF motifs were identified. The distance of a SNP from the nearest TF footprint is defined as the distance between the SNP to the nearest end of the corresponding footprint. The average size of the footprint is 12 bp. **(C)-(G)** Distribution of ASoC SNP distance from their nearest TF footprint in each cell type. The distance (bp) is expressed in $\log_{10}$ scale (x-axis) in (B) to (G).

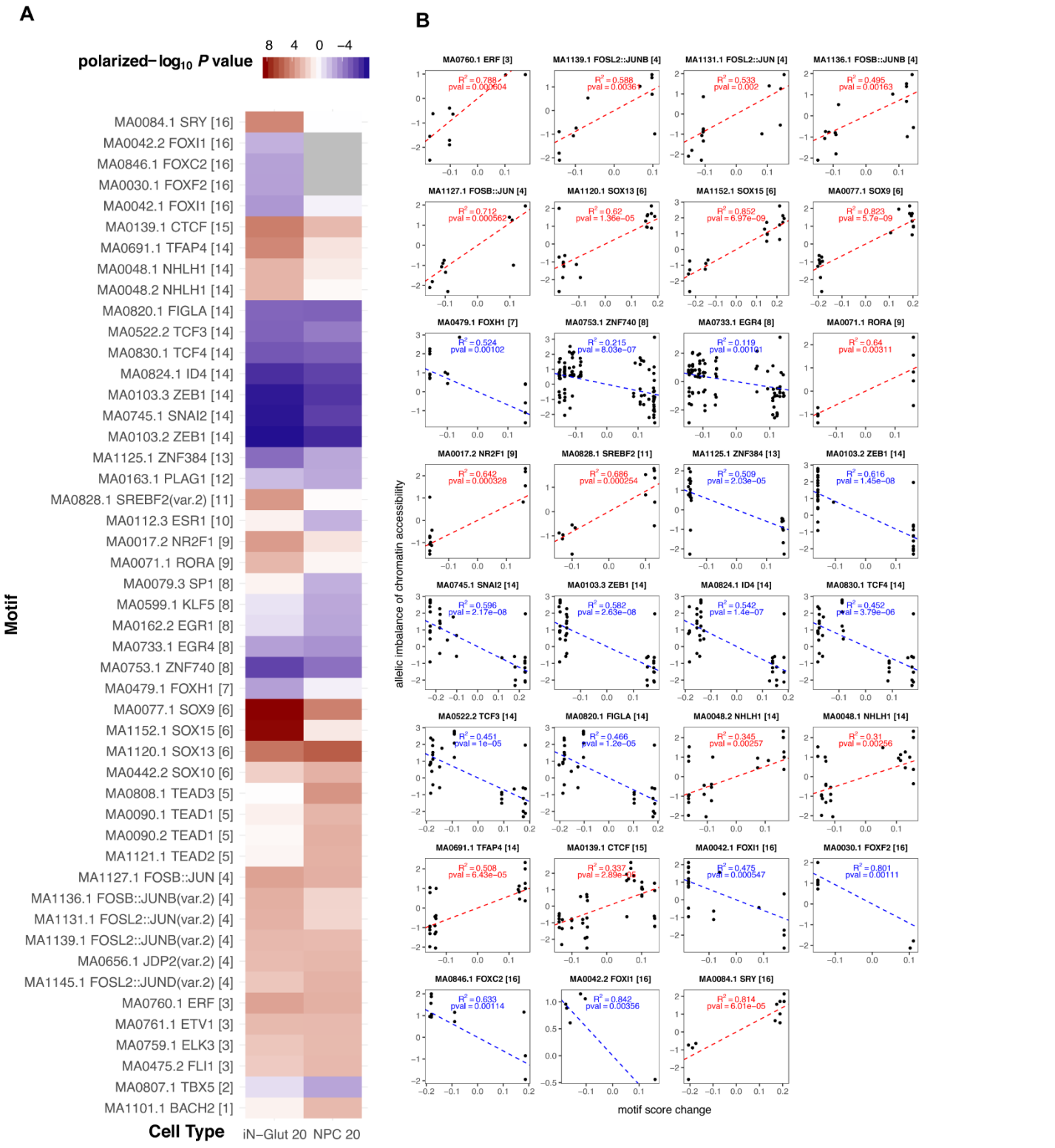
A
polarized −log10 P value
8 4 0 −4
Motif
MA0084.1 SRY [16]
MA0042.2 FOXI1 [16]
MA0846.1 FOXC2 [16]
MA0030.1 FOXF2 [16]
MA0042.1 FOXI1 [16]
MA0139.1 CTCF [15]
MA0691.1 TFAP4 [14]
MA0048.1 NHLH1 [14]
MA0048.2 NHLH1 [14]
MA0820.1 FIGLA [14]
MA0522.2 TCF3 [14]
MA0830.1 TCF4 [14]
MA0824.1 ID4 [14]
MA0103.3 ZEB1 [14]
MA0745.1 SNAI2 [14]
MA0103.2 ZEB1 [14]
MA1125.1 ZNF384 [13]
MA0163.1 PLAG1 [12]
MA0828.1 SREBF2(var.2) [11]
MA0112.3 ESR1 [10]
MA0017.2 NR2F1 [9]
MA0071.1 RORA [9]
MA0079.3 SP1 [8]
MA0599.1 KLF5 [8]
MA0162.2 EGR1 [8]
MA0733.1 EGR4 [8]
MA0753.1 ZNF740 [8]
MA0479.1 FOXH1 [7]
MA0077.1 SOX9 [6]
MA1152.1 SOX15 [6]
MA1120.1 SOX13 [6]
MA0442.2 SOX10 [6]
MA0808.1 TEAD3 [5]
MA0090.1 TEAD1 [5]
MA0090.2 TEAD1 [5]
MA1121.1 TEAD2 [5]
MA1127.1 FOSB::JUN [4]
MA1136.1 FOSB::JUNB(var.2) [4]
MA1131.1 FOSL2::JUN(var.2) [4]
MA1139.1 FOSL2::JUNB(var.2) [4]
MA0656.1 JDP2(var.2) [4]
MA1145.1 FOSL2::JUND(var.2) [4]
MA0760.1 ERF [3]
MA0761.1 ETV1 [3]
MA0759.1 ELK3 [3]
MA0475.2 FLI1 [3]
MA0807.1 TBX5 [2]
MA1101.1 BACH2 [1]
Cell Type
iN−Glut 20
NPC 20
B
MA0760.1 ERF [3]
MA1139.1 FOSL2::JUNB [4]
MA1131.1 FOSL2::JUN [4]
MA1136.1 FOSB::JUNB [4]
MA1127.1 FOSB::JUN [4]
MA1120.1 SOX13 [6]
MA1152.1 SOX15 [6]
MA0077.1 SOX9 [6]
MA0479.1 FOXH1 [7]
MA0753.1 ZNF740 [8]
MA0733.1 EGR4 [8]
MA0071.1 RORA [9]
MA0017.2 NR2F1 [9]
MA0828.1 SREBF2 [11]
MA1125.1 ZNF384 [13]
MA0103.2 ZEB1 [14]
MA0745.1 SNAI2 [14]
MA0103.3 ZEB1 [14]
MA0824.1 ID4 [14]
MA0830.1 TCF4 [14]
MA0522.2 TCF3 [14]
MA0820.1 FIGLA [14]
MA0048.2 NHLH1 [14]
MA0048.1 NHLH1 [14]
MA0691.1 TFAP4 [14]
MA0139.1 CTCF [15]
MA0042.1 FOXI1 [16]
MA0030.1 FOXF2 [16]
MA0846.1 FOXC2 [16]
MA0042.2 FOXI1 [16]
MA0084.1 SRY [16]
allelic imbalance of chromatin accessibility
motif score change

Fig. S16: Correlation analysis of the predicted TF motif disruption scores with the ASoC SNP allelic imbalance of chromatin accessibility.

(A) Heatmap of correlations of motif disruption scores vs. allelic imbalance of chromatin accessibility for all 48 TFs that show significant correlations (linear regression, FDR<0.05) in either iN-Glut-20 (n=31 TFs) or NPC-20 (n=32 TFs). Correlation levels are shown as polarized -$\log_{10}$ of regression *p*-values, indicating the significance and direction of association. **(B)** All TFs in iN-Glut-20 that exhibit significant correlation (FDR < 0.05) between SNP effect on motif match (x-axis) and on allelic chromatin accessibility (y-axis, $\log_2$ of alternative / reference allele read counts). Note both positive (red) and negative (blue) correlations are observed for TFs as putative activators or repressors, respectively. For both (A) and (B), the numbers in the brackets represent the cluster each motif belongs, based on the similarity of their position weight matrices.

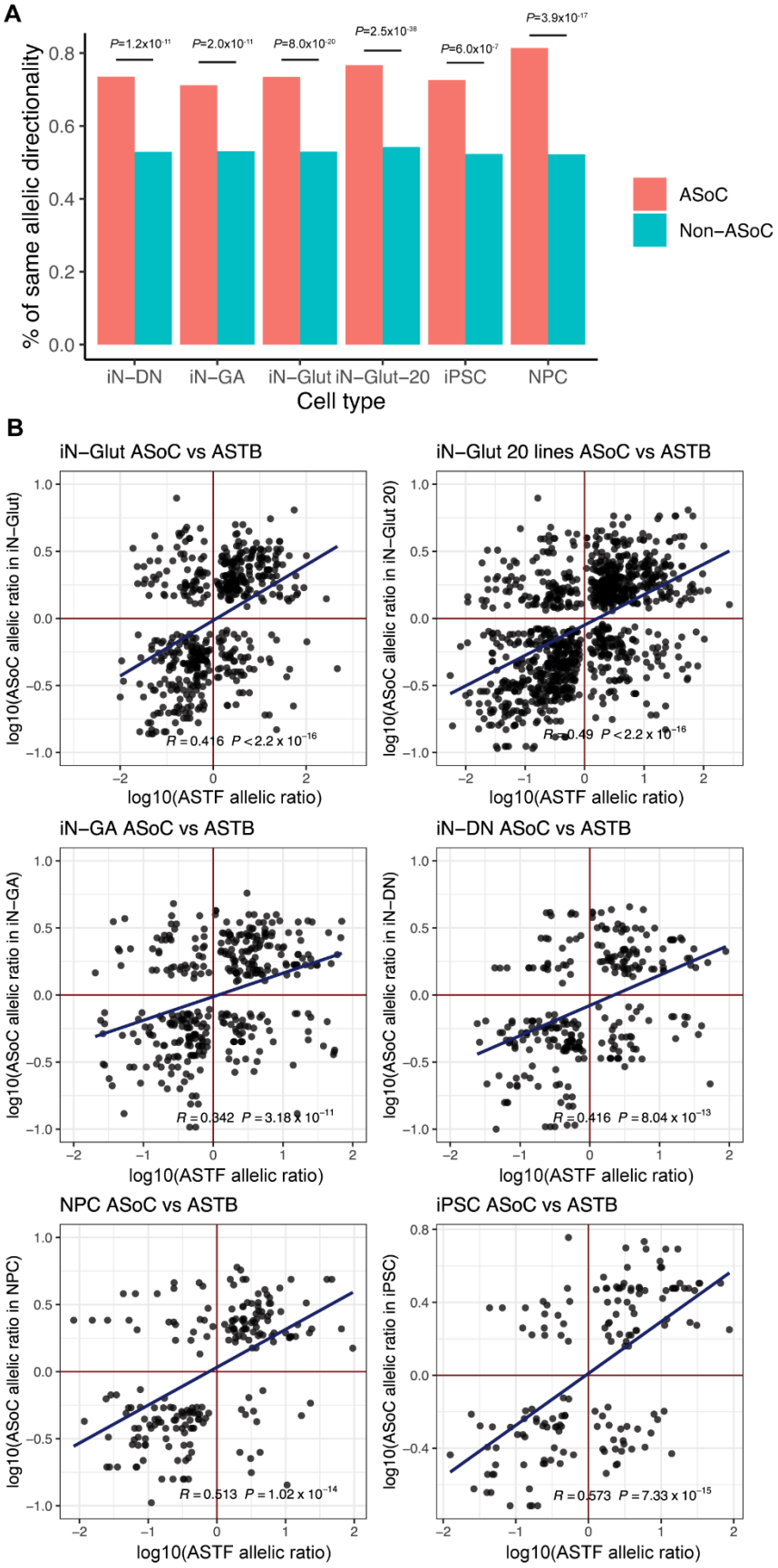
A
P=1.2x10-11
P=2.0x10-11
P=8.0x10-20
P=2.5x10-38
P=6.0x10-7
P=3.9x10-17
% of same allelic directionality
0.0
0.2
0.4
0.6
0.8
iN–DN
iN–GA
iN–Glut
iN–Glut–20
iPSC
NPC
Cell type
ASoC
Non–ASoC
B
iN–Glut ASoC vs ASTB
log10(ASoC allelic ratio in iN–Glut)
R = 0.416 P < 2.2 x 10−16
log10(ASTF allelic ratio)
iN–Glut 20 lines ASoC vs ASTB
log10(ASoC allelic ratio in iN–Glut 20)
R = 0.49 P < 2.2 x 10−16
log10(ASTF allelic ratio)
iN–GA ASoC vs ASTB
log10(ASoC allelic ratio in iN–GA)
R = 0.342 P = 3.18 x 10−11
log10(ASTF allelic ratio)
iN–DN ASoC vs ASTB
log10(ASoC allelic ratio in iN–DN)
R = 0.416 P = 8.04 x 10−13
log10(ASTF allelic ratio)
NPC ASoC vs ASTB
log10(ASoC allelic ratio in NPC)
R = 0.513 P = 1.02 x 10−14
log10(ASTF allelic ratio)
iPSC ASoC vs ASTB
log10(ASoC allelic ratio in iPSC)
R = 0.573 P = 7.33 x 10−15
log10(ASTF allelic ratio)

Fig. S17: Comparison of the allelic effects of ASoC SNPs on chromatin accessibility and on transcription factor binding.

(A) Percent of SNPs whose directions of allelic effects on chromatin accessibility in iPSC or neuronal cell types agree with the directions of their allelic effects on TF-binding (from ChIP-seq data). Our ASoC SNPs (FDR < 0.1, binomial test) and non-ASoC SNPs in each cell type were intersected with a list of SNPs that show allele-specific transcription factor binding (ASTF) (FDR < 0.05, binomial test) identified from a large collection of TF ChIP-seq data across multiple cell types (mainly from ENCODE) (*19*) (Table S16). *P*-values show the significance of enrichment of SNPs with matched directions in ASoC vs. non-ASoC SNPs (Fisher's exact test). **(B)** Pearson's correlation of the allelic ratios of ATAC-seq reads (Y-axis) of ASoC SNPs and their corresponding allelic ratios of ASTF ChIP-seq reads (X-axis).


A
−log10 enrichment P value
0 5 10 15 20
TF expression (normalized log2RPKM)
−5 0 5
iPSC
NPC
iN−Glut
iN−DN
iN−GA
ATF1 ATF3 ATOH1 BATF BRN2 CTCF ELK4 JUNB LHX2 NEUROG2 OCT4−SOX2−TCF−NANOG PAX6 RFX2 RFX5 SOX2 TCF4 TEAD2
B
iPSC
−log10 P
%
CTCF BORIS OCT4−SOX2−TCF−NANOG SOX3 OCT4 SOX21 OCT6 BRN1 TEAD1 SOX10 ZIC3 TEAD3 TEAD4 UNKNOWN−ESC−ELEMENT SOX15 SOX2 TEAD ZIC2 OCT11 ZIC SOX6 SOX4 TEAD2 SOX17 SP5 OCT2 SOX9 KLF1 SP1 SOX7
NPC
STAT5 STAT6 JUN−AP1 FOSL2 JUNB FOS FRA2 FRA1 SOX10 BATF ATF3 SOX21 AP−1 STAT4 SOX9 SOX3 BACH2 SOX2 TEAD3 TEAD1 CHR SMAD2 TEAD4 SOX15 SOX17 ZFX SOX6 BACH1 SOX4 SOX7
iN−Glut
STAT5 SOX10 SOX21 SOX9 STAT6 CTCF SOX3 SOX2 SMAD2 STAT4 CHR ATOH1 SMAD3 BHLHA15 SOX15 AP4 BMYB SCL NEUROD1 SOX6 NUR77 TWIST2 SOX17 ZFX TCF4 TCF21 NEUROG2 SOX4 PBX3 EWS:ERG−FUSION
iN−DN
NF1 RFX2 FRA1 FOS FRA2 FOSL2 RFX1 JUNB RFX JUN−AP1 ATF3 BATF RFX5 X−BOX AP−1 TEAD1 NF1−HALFSITE TLX? TEAD3 STAT5 RFX6 TEAD4 FOX:EBOX STAT4 STAT6 BACH2 TEAD FOXA2 FOXM1 FOXA1
iN−GA
EN1 NKX6.1 ATOH1 DLX2 LHX9 LHX2 LHX3 BHLHA15 TWIST2 NEUROG2 DLX1 NEUROD1 TCF4 TCF21 AP4 DLX5 ISL1 LHX1 ASCL1 CTCF TCF12 OLIG2 NANOG DLX3 SCL MYF5 MYOD MYOG RFX2 RFX

Fig. S18: Cell-type-specfic enrichment of TF-binding motif in ASoC SNPs and OCRs.

(A) Bubble plot shows the enrichment (-$\log_{10}P$ value, color scale) of specific TF motifs intersecting ASoC SNP sites (±50 bp) and the corresponding TF's expression level (normalized $\log_2$RPKM, circle size). All ASoC SNPs from the core-8 lines in each cell type were used for enrichment analysis by using HOMER (*49*). **(B)** Bar plots show the enrichments of top 25 TF motifs in OCRs (top panel) of each cell type and the percentage of OCRs (bottom panel) containing the top 25 enriched TF motifs. All peaks called by MACS2 with FDR < 0.05 from each core-8 set of each cell type were used for enrichment analysis. The enrichment *P* value (-$\log_{10}P$) of TF-binding motifs and the percentage of OCRs in each cell type were calculated by HOMER (*49*). The genomic backgrounds used for enrichment analyses are randomly sampled genomic sequences per HOMER default.

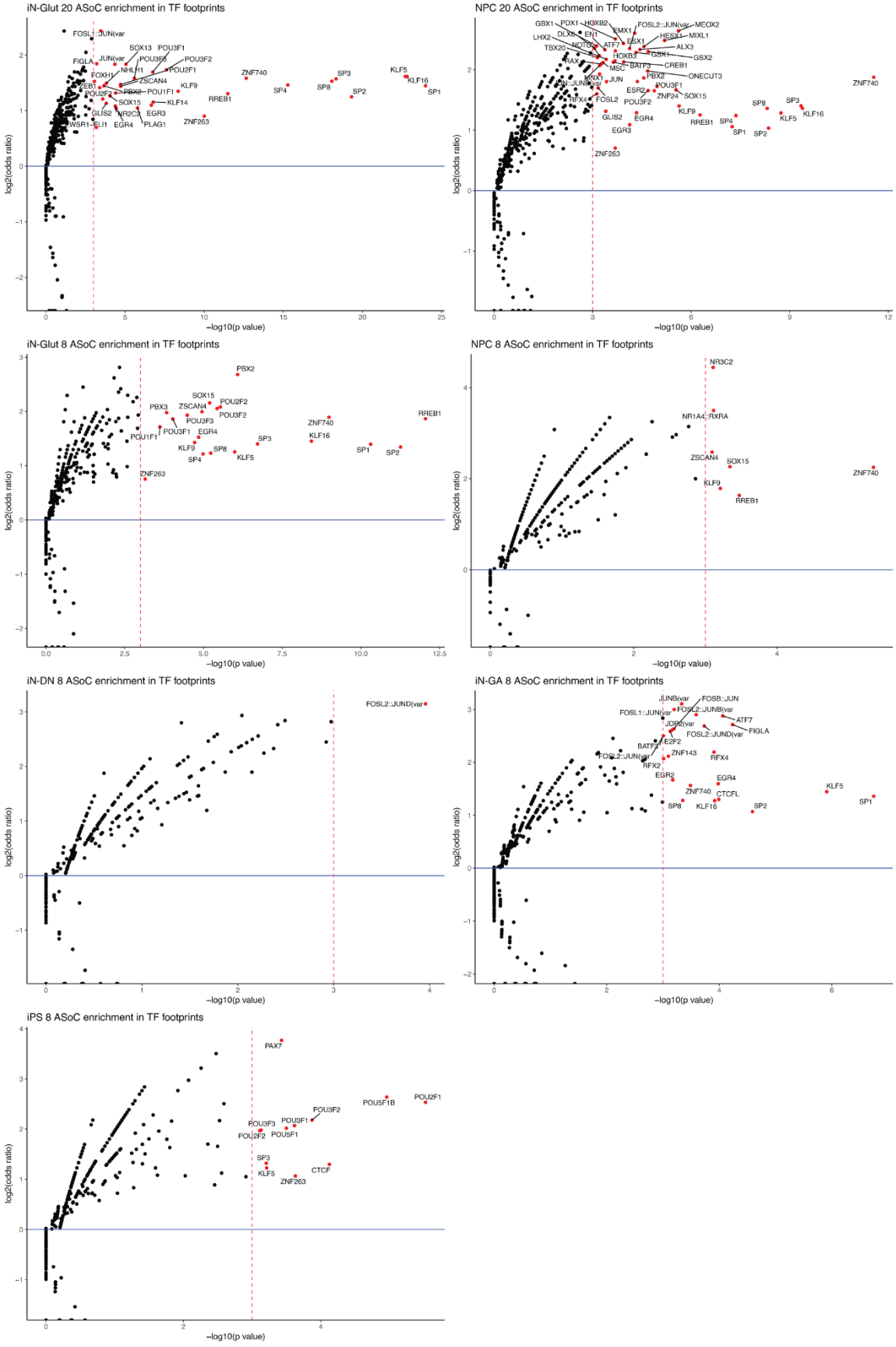
iN-Glut 20 ASoC enrichment in TF footprints
NPC 20 ASoC enrichment in TF footprints
iN-Glut 8 ASoC enrichment in TF footprints
NPC 8 ASoC enrichment in TF footprints
iN-DN 8 ASoC enrichment in TF footprints
iN-GA 8 ASoC enrichment in TF footprints
iPS 8 ASoC enrichment in TF footprints
log2(odds ratio)
−log10(p value)

Fig. S19: Cell-type-specific enrichment of TF-binding footprints in ASoC SNPs.

For each TF, we counted the number of ASoC SNPs that are located in the footprints identified in the given cell type and compared this with expectation (vs. non-ASoC SNPs in OCRs) using Fisher's exact test. The $-\log_{10}$ value of test *p*-values were plotted against $\log_2$ of the test odds ratios, and the vertical red line corresponds to a *p*-value threshold of 10^{-3}.

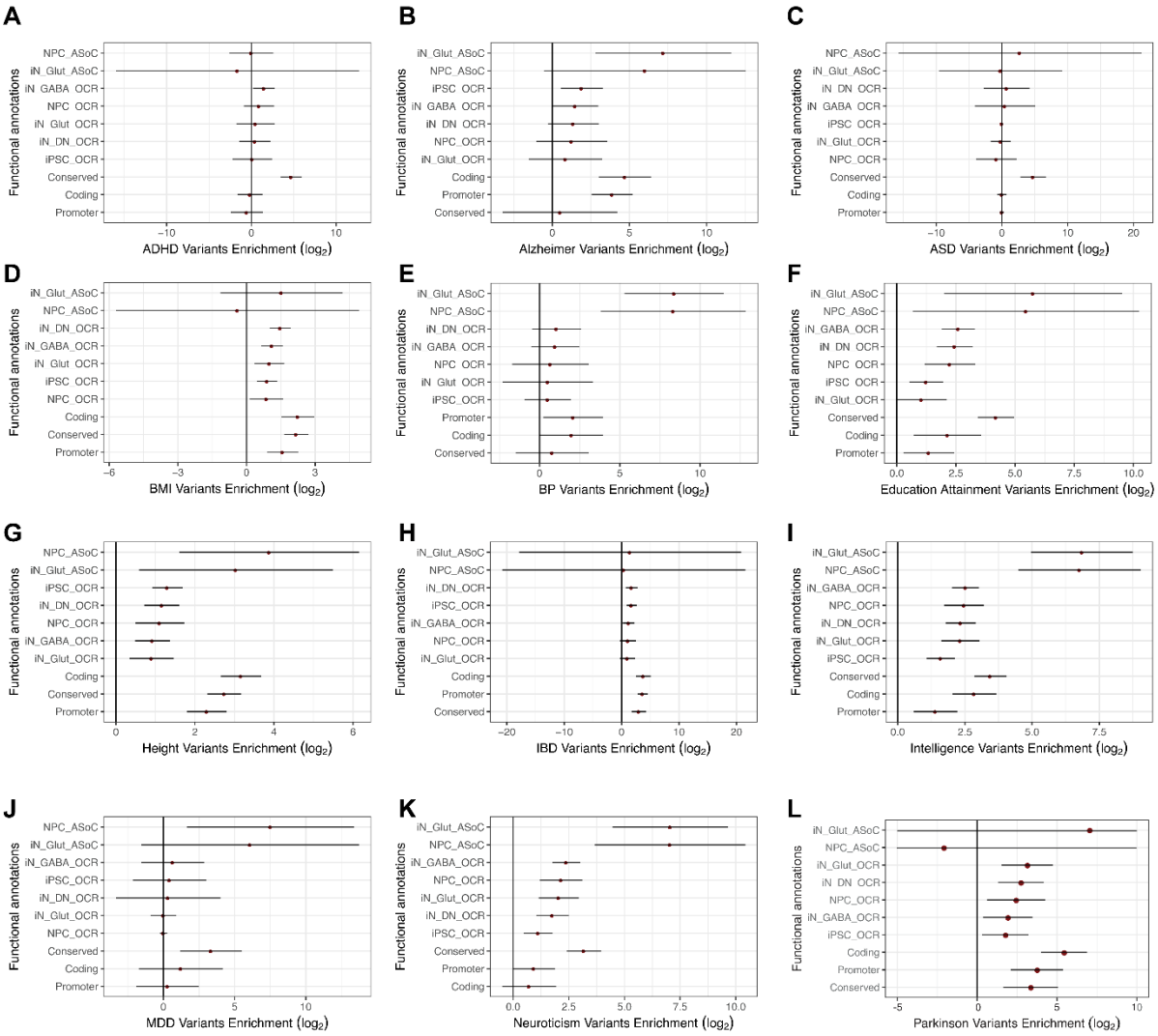

Fig. S20: Enrichment of ASoC SNPs, OCRs, and annotated genomic regions for GWAS risk variants of different diseases and traits.

(A-L): Enrichment analysis results from TORUS. Red dots indicate the $\log_2$ odds ratio of enrichment of ASoC SNPs (vs genomic non-ASoC SNPs), and bar ranges represent 95% confidence interval. Diseases/traits that did not show enrichment in OCR or ASoC in Figure 3C are not shown.

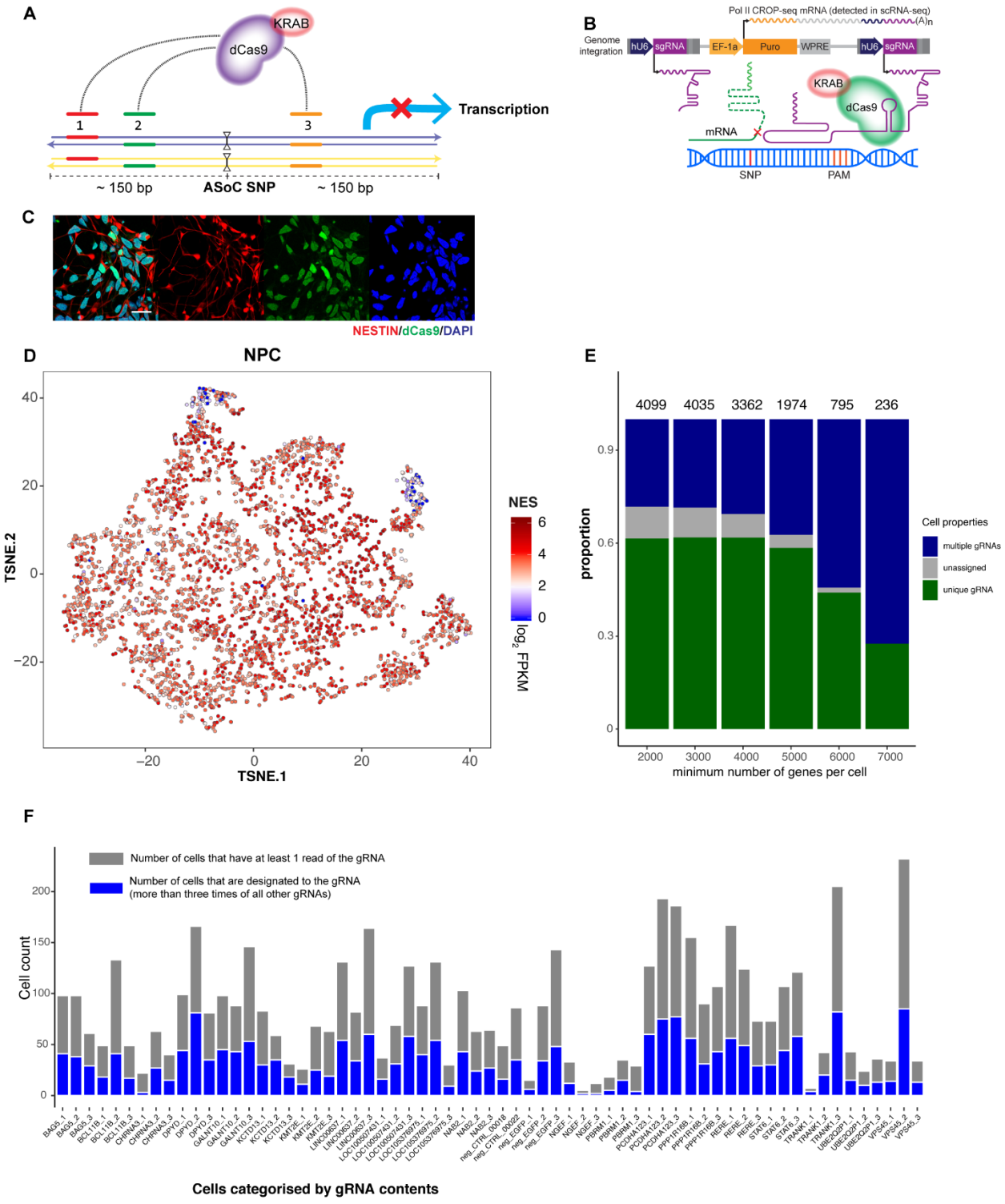

Fig. S21: The design of a modified CROP-seq approach to screen for cis-target genes of 20 selected ASoC SNP sites.

(A) Schematics of gRNA design. For each of the 20 ASoC SNP sites, we designed three targeting guide RNAs that flank within 150 bp of the targeted-SNP site. The dCas9-KRAB-gRNA complex was expected to repress the transcriptional activity near the targeted-SNP locus. **(B)** Schematics of

our modified CROP-Seq expression construct adapted from (*20*). The construct was customized so that, upon genome integration, the sgRNA cassette downstream of human U6 promoter would get activated and produce sequence-specific gRNA to form dCas9-KRAB-gRNA complex near the targeted SNP site and block transcription. Also, a second cassette harboring the puromycin-resistant gene would be transcribed independently under the regulation of an EF-1a promoter. The second transcript also carried the gRNA sequence as well as the poly-A tail that allows capture and detection of specific gRNAs during single-cell RNA-seq (see Methods). **(C)** Immunofluorescence images confirm the stable expression of dCas9 (fused with KRAB) in NPCs that were used for CROP-seq screening. Green: dCas9 protein; red: *NESTIN* (*NES*), an NPC marker; blue: DAPI. **(D)** T-distributed Stochastic Neighbor Embedding (T-SNE) plot of the 2,522 NPCs assigned with unique gRNAs. Note the homogenous population of NPC expressing the NPC-specific *NES.* **(E)** The proportion of single cells that contained single unique gRNAs as a function of the minimum number of expressed genes/cell. Note that >60% of the single cells were assigned with single unique gRNAs and were used in our CROP-seq analysis. **(F)** Bar graph showing the number of cells (Y-axis) that express a specific gRNA (X-axis). their gRNA contents. Blue bar: cells were assigned with a unique gRNA and used for differential gene expression analysis. Five negative control gRNAs were included.

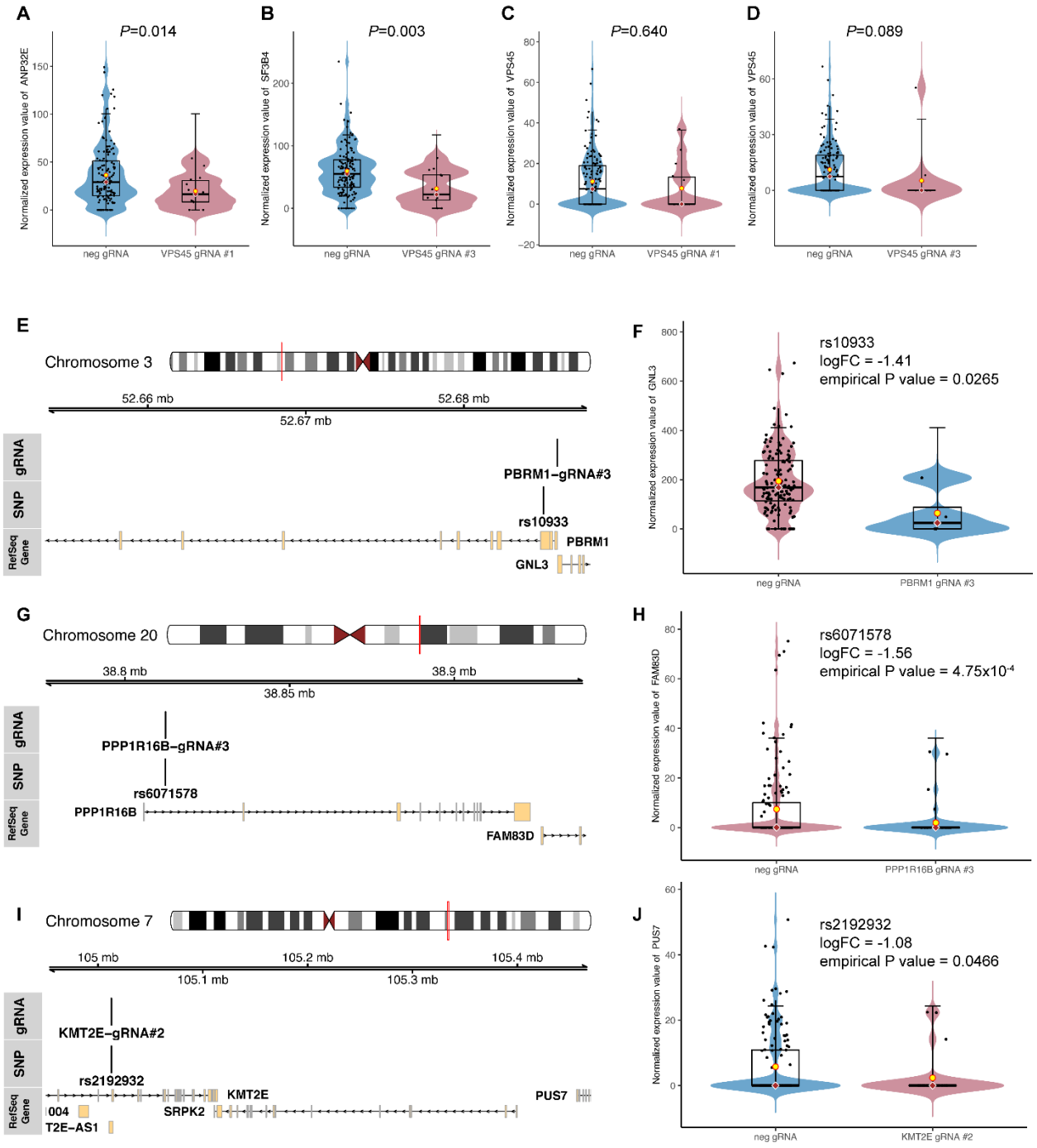

Fig. S22: Cis-target genes identified in CROP-seq screening in NPCs for some ASoC SNP site.

(A-B) Violin plots showing the reduced expression of *ANP32E* and *SF3B4* in single NPCs containing *VPS45* gRNA_3 (vs. control gRNAs) targeting to rs2027349, respectively. **(C-D)** Violin plots showing the reduced expression of *VPS45* in single NPCs containing *VPS45* gRNA #1 (C) and gRNA #3 (D) targeting to rs2027349. **(E-J)** ASoC SNP sites alters expression of a distal (rather than the nearest gene or >10kb) gene in NPCs. Left panels: local genomic map of each ASoC SNP site showing the genes around the ASoC SNP and the targeting site of the

gRNA. Right panels: the corresponding violin plots showing the reduced expression of a distal gene (*GNL3*, *FAM83D* and *PUS7*, respectively) of an ASoC SNP sequence in single NPCs that contain the specific SNP-targeting gRNA (vs. control gRNAs). Error bar: 95% CI; box: median ±25%; yellow oval with red stroke: mean expression level; red oval with white stroke: median expression level. Differentia expression in scRNA-seq was analyzed by EdgeR, and the empirical *p*-value was derived from permutation (see Method).

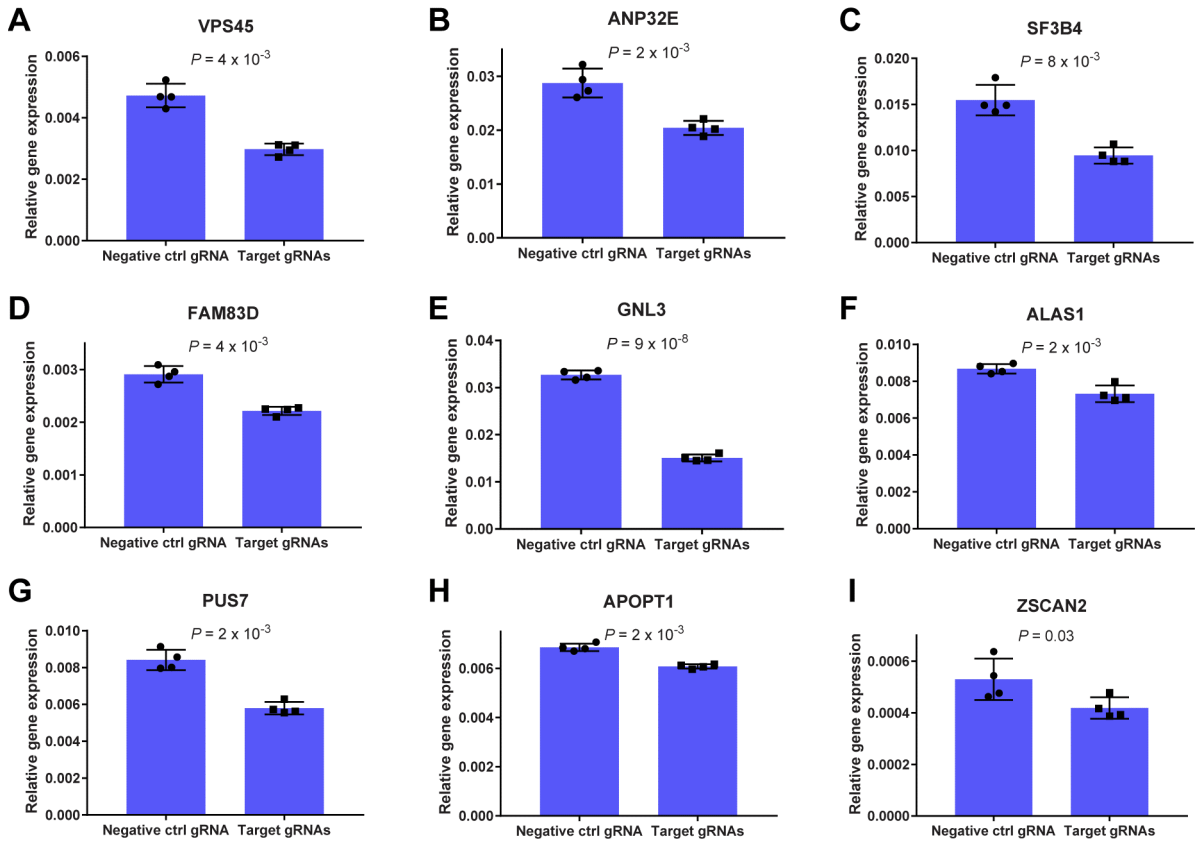

Fig. S23: CRISPRi followed by qPCR validation of the 9 selected cis-target genes identified by CROP-seq.

(A-I) Bar graphs showing differential expression of a CROP-seq nominated *cis*-target gene between NPCs containing a specific gRNA targeting to the ASoC SNP site corresponding to the *cis*-target gene. Similar to the CROP-Seq screen, the CRISPRi/qPCR validation was carried out in NPCs stably expressing dCas9/KRAB using a mixture of gRNAs that target to the ASoC SNP sites corresponding to the assayed *cis*-target gene. Negative control gRNA group contained 3 gRNAs targeting to EGFP. Relative gene expression (Y-axis; measured by qPCR) was normalized to *GAPDH*. Each dot on the bar chat represents a biological replicate (*i.e.*, an independent cell culture; n=4). *P* values were derived from unpaired Student's *t*-test with Welch's correction. Error bar: standard deviation.

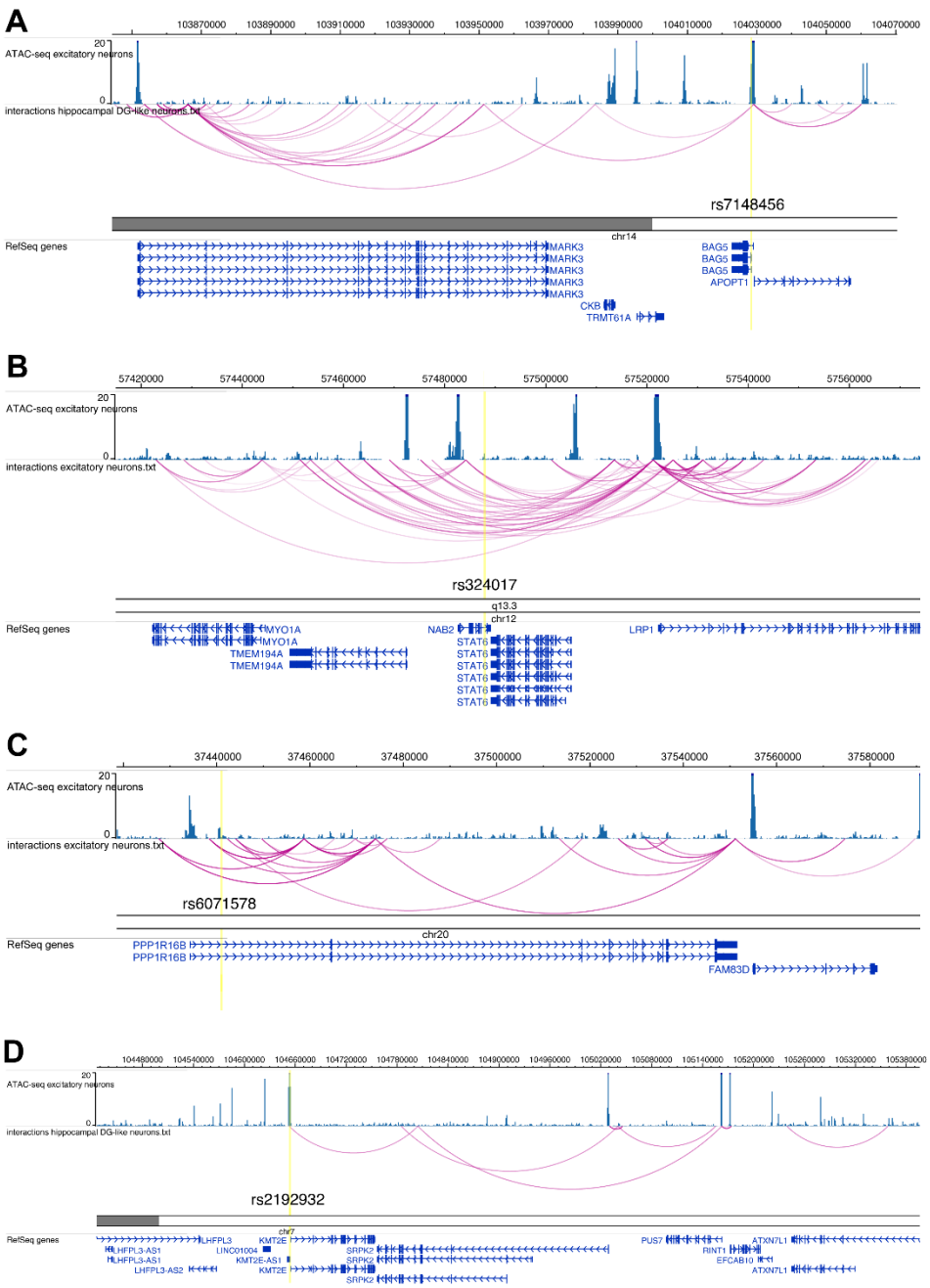

Fig. S24: Regional brain/neural Hi-C contacts for ASoC sequences identified to have CROP-seq distal (rather than the nearest gene or >10kb) *cis*-genes.

Cis-genes of an ASoC sequence in CROP-seq was defined as genes showing differential expression (empirical $P < 0.05$, by permutation test) in target gRNA group vs. control negative gRNA group of cells within 500 kb interval (+/- 500kb of the ASoC SNP site). Hi-C chromatin contact in iPSC-derived excitatory neurons and/or brain hippocampus neurons (*18*) were directly graphed from legacy WashU Epigenome Browser (session ID in parentheses):

epigenomegateway.wustl.edu/legacy/?genome=g19&session=8OCs2rkpEA (brain_pchic_nature_genetics_00). ASoC SNP position is highlighted with a vertical yellow line. **(A)** *BAG5* (rs7148456) locus with distal CROP-seq target *APOPT1*, a shared target with excitatory neuron Hi-C data. **(B)** *NAB2* (rs324017) locus with distal CROP-seq targets *LRP1*,*PTGES3* and *MARS*, of which *LRP1* is a shared target of excitatory neuron Hi-C data. **(C)** *PPP1R16B* (rs6071578) locus with distal CROP-seq target *FAM83D*. **(D)** *KMT2E* (rs2192932) locus with distal CROP-seq target *PUS7* (hippocampus Hi-C data was shown since there is no Hi-C contact at the SNP site in excitatory neuron). Although *PUS7* is not a direct Hi-C target, it was confirmed by our independent CRISPRi/qPCR validation (Fig. S23I). All genomic positions are based on hg19 in this figure due to the data source limitation.

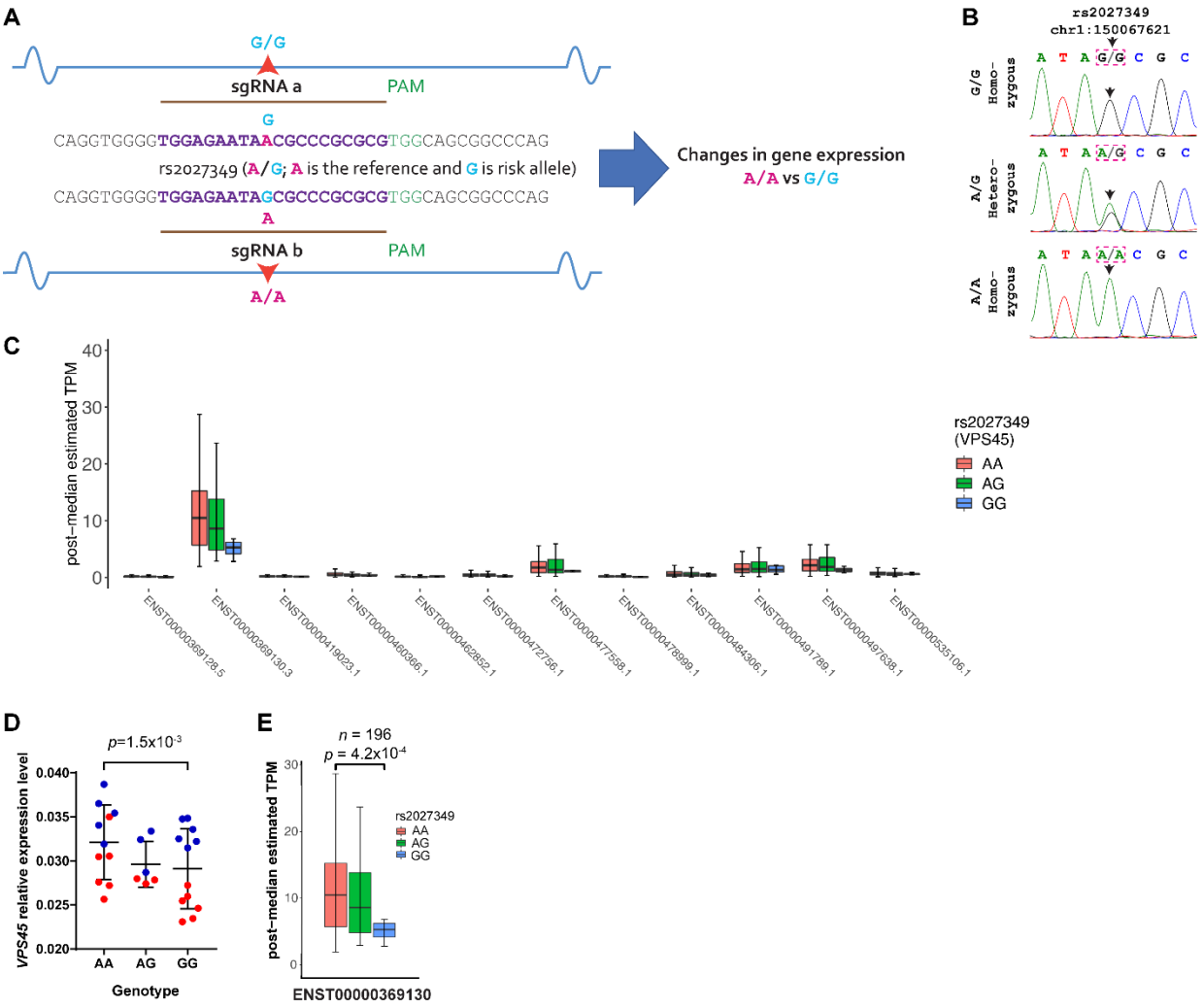

Fig. S25: CRISPR/Cas9-editing of the ASoC SNP rs2027349 at the VPS45 locus.

(A) Schematic of CRISPR/Cas9 base editing at SNP rs2027349. Two iPSC lines from different donors heterozygous (A/G) for rs2027349 were edited to homozygous A/A or homozygous G/G at the SNP site. Three different CRISPR-edited iPSC clones were generated for each genotype. **(B)** Confirmation of SNP editing results by Sanger sequencing CRISPR-edited iPSC clones. Arrows point to the SNP site. **(C)** Whisker plot of the expression level (in TPM, transcript per million; Y-axis) of different *VPS45* transcript isoforms (RNA-seq data) in post-mortem brains of healthy Caucasian individuals (n = 196; from CommonMind). Note the major transcript isoform (ENST00000369130) with high expression. The differential expression of ENST00000369130 between individuals with different genotypes is shown separately in (E). **(D)** qPCR result of the relative expression level of *VPS45* (all transcript isoforms) in NPCs derived from CRISPR-edited isogenic iPSC lines with different genotypes (AA, AG, and GG) at SNP rs2027349. The genotypic trend effect on the expression of *VPS45* was first analyzed for the two lines (denoted by red or blue color) separately (β = -0.0022, s.e. = 0.000067, and p=0.005 for line #1; β = -0.001, s.e. = 0.00007, and p=0.189 for line #2; linear regression test). The Fisher's test combining the two lines gave p=0.0015, with A allele associated with higher *VPS45* expression (n = 2-3 clones/line each with 2-3 independent experiments for AA or GG). s.e., standard error. **(E)** *VPS45* expression

(ENST00000369130) in CommonMind post-mortem brain RNA-seq samples of Caucasian control individuals (n = 196), showing A allele is associated with higher *VPS45* expression. Unpaired Student's t-test with Welch's correction was used.

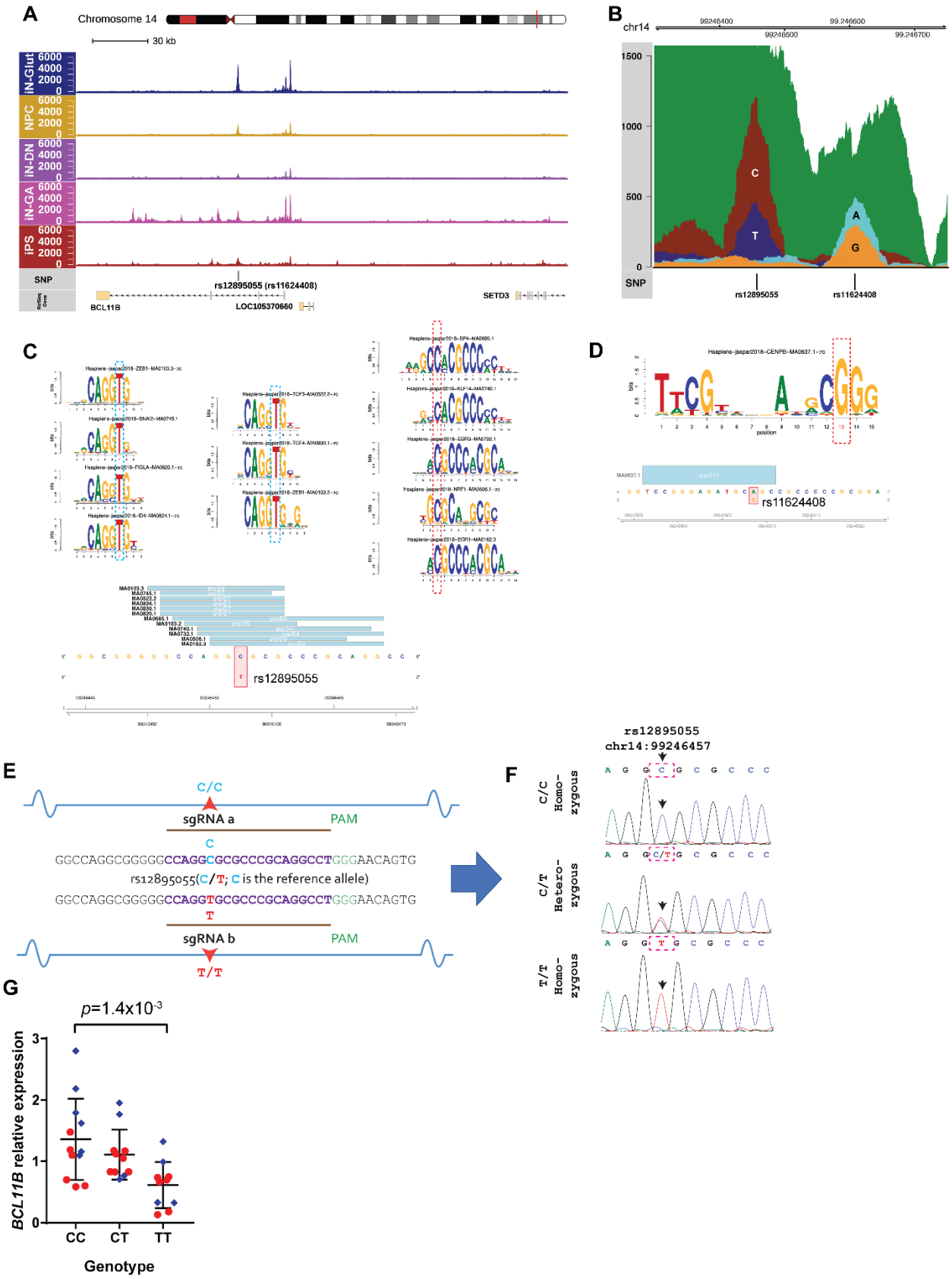
A
Chromosome 14
30 kb
iN-Glut
NPC
iN-DN
iN-GA
iPS
SNP
rs12895055 (rs11624408)
BCL11B
LOC105370660
SETD3
B
chr14
99246400
99246500
99246600
99246700
C
T
A
G
rs12895055
rs11624408
C
rs12895055
D
rs11624408
E
C/C
sgRNA a
PAM
GGCCAGGCGGGGGCCAGGCGCGCCCGCAGGCCTGGGAACAGTG
rs12895055(C/T; C is the reference allele)
GGCCAGGCGGGGGCCAGGTGCGCCCGCAGGCCTGGGAACAGTG
T
sgRNA b
PAM
T/T
F
rs12895055
chr14:99246457
C/C Homo-zygous
C/T Hetero-zygous
T/T Homo-zygous
G
p=1.4x10-3
BCL11B relative expression
CC
CT
TT
Genotype

Fig. S26: Cas9-editing of the ASoC variant rs12895055 at the BCL11B locus.

(A) ATAC-seq OCR peaks at the *BCL11B* locus in different cell types. Note the iN-Glut specific OCR that flanks the ASoC SNP in iN-Glut. **(B)** ATAC-seq reads pile-up showing allelic bias at SNP sites rs12895055 and another linked SNP rs11624408 (*i.e.*, both show ASoC; n = 6 lines) in iN-Glut cells (heterozygous)**.** Note the more robust ASoC for rs12895055 compared to rs11624408. **(C-D)** TF-binding motif disruption at SNP sites rs12895055 and rs11624408, respectively. Sequence logos of all the TF motifs were generated in motifbreakR. The circled positions in the motif logos correspond to where the ASoC SNP position. A larger bit number (Y-axis) in a motif logo indicates a more conserved nucleotide position. **(E)** Schematic of CRISPR/Cas9 editing of SNP rs12895055. Two different iPSC lines from two donors heterozygous (C/T) for rs12895055 were edited to homozygous C/C or homozygous T/T at the SNP site. 2-3 different CRISPR-edited iPSC clones were generated for each genotype. **(F)** Confirmation of rs12895055 editing results by Sanger sequencing in CRISPR-edited iPSC clones. Arrows point to the SNP site**. (G)** qPCR results of the relative expression level of *BCL11B* in iN-Glut neurons derived from CRISPR-edited isogenic iPSC lines with different genotypes (CC, CT and TT) at SNP rs12895055. The genotypic trend effect on the expression of *BCL11B* was first analyzed for the two lines (denoted by red or blue color) separately (β = -0.21, s.e. = 0.008, and p=0.03 for line #1; β = -0.52, s.e. = 0.19, and p=0.018 for line #2; linear regression test). The Fisher's test combining the two lines gave p=0.0014, with T allele associated with reduced expression (n = 2-3 clones/line each with 2-3 independent experiments for each genotype). s.e., standard error.

Supplementary Table Legends

Table S1. Subjects used for generating iPSC lines

Table S2: Next-gen sequencing statistics of ATAC-Seq and RNA-Seq results.

Table S3. OCR peaks overlap with PsychENCODE brain ATAC-peaks.

To make the peak definition comparable to PsychENCODE, OCR peak was defined as summit +/- 250bp. Peaks overlap more than 25% were defined as overlapping peaks. Total number of PsychENCODE peaks: 117,935.

Table S4: Annotation of ASoC SNPs in iPSC samples.

Table S5: Annotation of ASoC SNPs in NPC samples.

Table S6: Annotation of ASoC SNPs in iN-Glut samples.

Table S7: Annotation of ASoC SNPs in iN-GA samples.

Table S8: Annotation of ASoC SNPs in iN-DN samples.

Table S9: Statistics of heterozygous and ASoC SNP numbers and percentage in different cell lines.

Table S10: ASoC SNPs called from 20 NPC samples.

Table S11: ASoC SNPs called from iN-Glut-20 and brain QTL annotation.

(p-values from QTL test were listed; #N/A = not available).

Table S12: Comparison between a public MPRA reporter gene assay dataset for a large number of SNPs (in LCL) (*15*) and our iN-Glut20 ASoC SNP dataset.

LogSkew.Comb =log2 allelic skew from combined LCL analysis (alternative/reference), C.Skew.logP=allelic skew p-value from combined LCL analysis, C.Skew.fdr=allelic skew fdr from combined LCL analysis. Only overlapping SNPs between the two datasets are listed here. FDR = False discovery rate.

Table S13: Comparison between a public MPRA reporter gene assay dataset for a large number of SNPs (in LCL) (*15*) and our NPC-20 ASoC SNP dataset.

LogSkew.Comb =log2 allelic skew from combined LCL analysis (alternative/reference), C.Skew.logP=allelic skew p-value from combined LCL analysis, C.Skew.fdr=allelic skew fdr from combined LCL analysis. Only overlapping SNPs between the two datasets are listed here. FDR = False discovery rate.

Table S14: Oligos designed for luciferase reporter gene assay of a selected set of ASoC SNPs in NPCs.

The assays for the reference allele and alternative allele are listed separately.

Table S15. ASoC SNP linkage to eQTL/HiC/promoters.

ASoC SNPs that can be resolved to putative cis-targets through linking to brain eQTL SNPs (CommonMind adult brain) (*17*), brain/neuronal HiC chromatin contacts (excitatory neurons or hippocampus neurons; Supplementary Table 2 in (*18*)) and/or promoter annotations (2 kb upstream/1 kb downstream of a TSS, transcripts of Gencode v31 comprehensive annotation set). Only eQTL SNPs within 500 kb of a promoter and with FDR < 0.05 were used in the comparison. For Hi-C data, as described in (*18*), interacting fragments were expanded to a minimum width of 5 kb if a fragment size is less than 5 kb.

Table S16: Comparison between a SNP dataset of allele-specific TF-binding (ASTB).

A Comparison between a SNP dataset of allele-specific TF-binding (ASTB) in various cell types (from ENCODE) (*19*) (Link of the dataset: https://www.biorxiv.org/content/10.1101/253427v1) and our ASoC datasets in iN-Glut20, and core 8 of iPSC, NPC, iN-Glut, iN-DN and iN-GABA. Only the overlapping SNPs between the two datasets are listed.

Table S17. Homer motif enrichment analysis for ASoC SNPs in each cell type.

Intervals used for enrichment test are SNP-flanking 50+/-50bp. HOMER known motifs were used.

Table S18. Homer motif enrichment analysis for OCRs (peaks) specific to each cell type.
The list of HOMER known motifs were used.

Table S19: ASoC SNPs and non-ASoC heterozygous SNPs that overlap with SZ genome-wide significant SNPs and credible SNPs in 20 iN-Glut lines.

Table S20: ASoC SNPs from 20 NPC lines.

Table S21: GWAS datasets that were used in TORUS enrichment analysis.

Table S22: SNP information and their corresponding gRNA sequences used for CROP-Seq multiplex.
Color code: red denotes lower value.

Table S23: Full differential expression analysis results for genes within 500kb of targeted SNPs in CROP-seq.
Table is sorted by "Empirical_pvalue" from small to large for each locus. "Locus_SNP": SNP name (dbSNP150) targeted by gRNA. "Locus_name": a custom name for targeted locus. "gRNA_name": gRNA names, for most loci we designed three gRNAs. "Cis Gene Name": genes within 500kb up or downstream of the targeted locus based on UCSC table browser (hg19 assembly). "logFC": log Fold change calculated by edgeR. "logCPM": log CPM (count per million) calculated by edgeR. "Emprical_Pvalue": p values calculated using permutations (see method for details). Yellow backgrounds denote genes that have empirical p value < 0.05 for the 6 loci highlighted in Table S25.

Table S24: qPCR assay ID and the sequences of qPCR primers (for customized assay).

Table 25: The 10 SZ-associated ASoC SNPs/sequences with CROP-seq cis-target genes and supporting evidence from brain eQTL, brain/neural HiC data.
Brain QTL data are from three different sources as shown as column headings. HiC chromatin contact evidence was also from three sources as indicated. The SNP-target associations were directly extracted from the Tables in those Hi-C papers. These 10 SZ ASoC SNPs were identified by CROP-seq to have cis-targets at permutation $p < 0.05$. Results from independent bulk CRISPRi/qPCR confirmation was also included. The 4 CROP-seq cis-targets (*VPS45*, *APOPT1*, *GNL3*, and *LRP1*) shared with brain eQTL and/or Hi-C data were bolded. "distal" = not adjacent gene and >2kb away, in red.

Table S26: Fine-mapping results for 20 loci. Fine-mapping results of 20 ASoC SNPs selected for CROP-seq experiments.
The PIP column shows the posterior inclusion probabilities of SNPs from fine-mapping analysis with Susie. Note Susie only reports PIPs for SNPs included in 95% credible set, so some SNPs have no PIP values, meaning these SNPs have low probabilities of being causal variants. The rest shows the annotations of the SNPs, with 1 means present and 0 absent. The six highlighted loci have PIP scores > 0.10 and at least one cis-gene nominated by CROP-seq. OCR: open chromatin region.

Table S27. Sequences of gRNA, repair oligos and on/off-target PCR primers used for rs2027349 editing (rs2027349 site is shown in red).

Table S28. Sequences of gRNA, repair oligos and on/off-target PCR primers used for CRISPR/Cas9 editing of BCL11B locus ASoC SNP rs12895055 (SNP site is shown in red).