Supplementary Materials

Table of Contents

Supplementary Methods

MPRA library cloning

Resuspended oligo pool (10pg/µl) was amplified using PrimeSTAR Max DNA Polymerase (Takara # R045A) with MPRA cloning forward primer and reverse primer to introduce the EcoRI and BamHI restriction sites upstream and downstream of the oligo, respectively. The amplified library was digested with EcoRI and BamHI, ligated into a pGreenFire vector (Addgene, #174103) with a blastocidin selection marker at a 1:2 vector:insert ratio, transformed into Stellar Competent cells (Takara #636763) in 40 parallel reactions, recovered overnight for 12 hours, and pooled. The expanded library was isolated by the Qiagen Plasmid Plus Maxi kit. We generated a miniP-miniLuc amplicated by PCR from plasmid pD2 (Addgene, #174105) digested both the miniP-miniLuc insert and the plasmid library with XhoI and XbaI (excising the filler region), ligating at a 1:8 vector:insert ratio, transformed into Stellar Competent cells (Takara # 636763) in parallel reactions, recovered overnight for 12 hours, pooled. This second step library isolated again by Qiagen Plasmid Plus Maxi kit. Multiple iterations of the cloning process were done and pooled to form the final plasmid library.

MPRA library virus generation

LentiX cells (passage < P8) were grown in 15cm plates until ~80% confluent. Plasmids pCMV R d8.91 (25 ug/plate), pUC-MDG VSVG (10 ug/plate), and the plasmid library (25 ug/plate) were transfected using Lipofectamine 2000 (Life Technologies). Supernatant was harvested 48 and 72 hours post transfection. GoStiX LentiX sticks were used to rapidly assess transfection efficacy (p24 reading >300). Supernatant was concentrated using LentiX concentrator (Takara) at a 3:1 vol:vol ratio of supernatant: concentrator, then aliquoted and frozen down to -80C.

General MPRA cell culture and infection

In each cell type, optimal blastocidin concentration was determined, and virus was titrated using CellTiterBlue assays to minimize virus toxicity and maximize infection efficiency. Additionally, average integrants per cell was determined for infected cells. Briefly, gDNA was extracted from infected cells post-selection via Qiagen tissue extraction kits. Serial dilutions of the original plasmid library and the gDNA were performed. qPCR was performed on all serial dilutions using primers designed for the oligo library sequences to determine the number of copies of the integrants present in each gDNA sample, using the formula: $\log_{10}$(copies) = PLASMID_INTERCEPT * C_q + PLASMID_SLOPE. Cell number for each gDNA sample was approximated based on the assumption that there is roughly 6.6 pg of gDNA per cell. The average integrants per cell was calculated by dividing the number of copies present in a gDNA sample by the number of estimated cells. Average number of integrants per cells greater than 4 were desired.

Infection and Culture of Astrocytes for Psych MPRA

Normal Human Astrocytes (Lonza, CC-2565) were cultured in Lonza Astrocyte Growth Media containing astrocyte basal medium, 7.5% FBS, ascorbic acid, rhEGF, GA-1000 (gentamicin sulfate-amphotericin), insulin, and L-glutamine. Astrocytes were seeded at a 5,000 cells/cm^2 density and maintained up to 70% confluence before splitting at 37C and 5% CO2. Growth medium was changed the day after seeding and every other day thereafter. For each biological replicate, roughly 1.5x10^7 cells were infected with the MPRA library using 8 ug/ml polybrene, selected for 24 hrs post-infection with 30 ug/ml blastocidin for 48 hrs, and grown until ~1.2x10^7 cells were collected, washed, and frozen at -80C.

Infection and Culture of Cell Lines for Psych MPRA

HEK293T (Takara, cat. no. 632180) cells were cultured in DMEM (Thermo Fisher Scientific cat. no. 11995065) supplemented with 10% FBS and 1X Penicillin-Streptomycin (Thermo Fisher Scientific cat. no. 15140122) at 37°C with 5% CO_2. Cell were lifted using 0.05% trypsin (Thermo Fisher Scientific cat. no. 25300054) and passaged 1:10 once they reached 80-90% confluence.

SH-SY5Y neuroblastoma cells (ATCC, CRL-2266) were grown in 1:1 F12 (Lonza, cat. no. 12-615F): EMEM (Thermo Fisher Scientific cat. no. 50-188-268FP). SH-SY5Y cells were differentiated in Neurobasal medium (Thermo Fisher no 21103049) with B27 (Thermo Fisher no 17504044) and Glutamax (Thermo Fisher no 35050061) supplements and 10 µM all-transretinoic acid (ATRA) according to the published protocol[1]. Half-media changes were made every day for 6 days.

IMR-32 neuroblastoma cells (ATCC, CCL-127) were grown in EMEM media. IMR-32 cells were differentiation by adding 1 mM dibutyryl-cAMP (Fisher: Stem Cell Technologies cat. no. 73884) and 2.5 µM BrdU (Fisher Scientific cat. no. B23151) in EMEM for 6 days. A full media change was done on day 3.

D283 medulloblastoma cells (ATCC, HTB-185) were grown in EMEM media in suspension. D341 medulloblastoma cells (ATCC, HTB-187) were grown in suspension in EMEM media supplemented with 20% FBS.

MPRA sample culture and processing for each cell line was done in parallel.

To perform the MPRA in cell lines, an MPRA lentiviral library titration was performed to determine the volume of concentrated lentivirus needed to achieve a high infection rate without negatively affecting cell growth. Titration was also performed to test the optimal concentration of blastocidin S HCl needed for a 48 hour selection. For each sample, 10 million cells were transduced with 700µL of the MPRA lentiviral library and 5 µg/mL polybrene. Cells were plated into two 10cm plates (5 million cells/plate) and cultured overnight. Approximately 24 hours post transduction, cells were lifted and plated into two 15cm plates in media supplemented with blasticidin S HCl. Non-transduced cells were fully selected within 48 hours. After 72 hours of drug selection, cells were collected and lysed in Qiagen buffer RLT Plus containing 2-mercaptoethanol. Transductions were performed in triplicate. Lysates were frozen at -80°C prior to performing RNA isolation using a Qiagen RNeasy Plus Mini Kit (Qiagen cat. no. 74136).

MPRAnalyze Model

SNVs with significant allele specific activity were determined per cell type using the R package MPRAnalyze[2] v1.4.0, which assumes a linear relationship between RNA and DNA (RNA = α DNA), where α represents a transcription rate. RNA counts (***r***) are approximated as a negative binomial distribution, and DNA counts (***d***) are fit to an underlying gamma distribution. We note that DNA plasmid library counts were used as baseline. Both DNA and RNA abundances are modelled via separate log-additive regression models. DNA is modelled with a design matrix (X_d) encoding a barcode-allele coefficient that assumes barcode and allelic (reference or alternate) effects independently contribute to oligo abundance. This allows a per-barcode and a per-allele estimation of DNA counts. RNA design matrix (X_r) includes a factor for the allele term only and was designed such that coefficients represent effects of the reference and alternate condition. Additionally, for controls we used sequences of both reference and alternate alleles (n=11 loci) that virtually no activity across all conditions, replicates, barcodes, and tissues as our background for transcriptional activity. In equation form, we represent the DNA model as:

$$log(\boldsymbol{d}) = X_d\,\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the DNA model coefficient and $X_d \sim$ barcode_allele.

The full RNA model is then:

$$log(\boldsymbol{r}) = log(\boldsymbol{d}) + log(\alpha)$$

$$log(\boldsymbol{r}) = X_d\,\boldsymbol{\beta} + log(\alpha) = X_d\,\boldsymbol{\beta} + X_r\boldsymbol{\gamma}$$

where $\boldsymbol{\gamma}$ is the RNA model coefficient and $X_r \sim$ allele specific co-efficient. We note that replicates were normalized prior to fitting to avoid batch-specific effects.

In words, we model the full model of the allele-specificity model to be:

RNA ~ replicate + allele

The reduced model (intercept-only only):

RNA ~ replicate

and models the null hypothesis which states that there is no allelic imbalance between the reference or alternate allelic condition.

Likelihood maximization was used to fit these full generalized linear models (GLM) to extract fitted coefficients for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and thereby α. α value corresponds to the transcriptional rate of the allelic element. Log2-fold change represents the log-transcriptional activity changes between alternative and reference allele sequences. A one-sided likelihood ratio test, within the MPRAnalyze package, is used to extract p-values by comparing the full model to the reduced model. P-values are multiple-hypothesis corrected using the R function p.adjust, method=fdr.

SNVs with allele specific activity were defined as those achieve a $|\ \log_2(\text{fold-change})| > 0.05$ and an FDR-corrected p-value < 0.05. Empiric p-values were plotted as a QQ-plot against expected p-values (given a uniform distribution between [0,1].

Likewise, a similar GLM model can be built assessing cell-type specificity where the full model:

RNA ~ replicate + allele + tissue + allele:tissue

Whereas the reduced model would be:

Batch-normed RNA ~ replicate + allele + tissue

This likelihood ratio test (FDR<0.05) was then used to assess the null hypothesis that tissue and allelic effects on RNA activity are independent from one another. This LRT was used to assess whether there were tissue-specific effects that were loci dependent.

Power Analysis

We used the simulateMPRA function in the MPRAnalyze package to perform a power analysis to determine how many barcodes are necessary for detection of a given level of fold change (allelic effect size). We generate the dispersion metrics for the simulated dataset by extracting fitted parameters from our own MPRA dataset. We tested power for at 4 different log2-fold change thresholds (1.2, 1.5, 2, 3), for a total of 20 variants, simulated 5 times with 5, 10, 20, 50, and 100 barcodes each. Results are shown in fig S1D.

Luciferase plasmid-based assays

401 bp fragments were designed for selected SNVs of interest, by extracting the genomic sequence (hg19) centered around the SNV of interest. Sequences were selected based on their linked eGenes of interest, MPRA significance at multiple time points, and magnitude of the alternate-to-reference log fold change. Luciferase Cloning adaptors were added upstream and downstream of the genomic instance, respectively. Sequences were synthesized as GeneBlocks (IDT). Fragments were cloned into a pGL4. 23 using In-Fusion Snap Assembly (TAKARA Cat 638943). Plasmid sequences were confirmed by Sanger sequencing and then transfected in SH-SY5Y cells with 4 number of replicates per sequence. Cells were harvested after 48 hours. Luciferase signal was measured using the Dual-Luciferase® Reporter Assay System (Promega) using Tecan Infinite M1000. Luciferase signal was calculated following manufacturing specifications. Briefly, firefly and Renilla blanks were subtracted for respective measurements. The firefly to Renilla ratio was calculated for all alternate and reference measurements, then normalized to the control empty vector ratio. Normalized ratios less than 1 were removed (n=10 daSNVs), as those sequences did not transfect well in the chosen cell system. p-values were calculated using the two-sided Mann-Whitney u-test.

Luciferase lentiviral-based assays

A lentiviral reporter construct was designed that contains a minimal promoter driving the expression of destabilized copGFP and luciferase separated by a T2A sequence. The construct also contains a CMV driven blasticidin S deaminase gene. Genomic sequences synthesized by IDT were inserted upstream of the minimal promoter by digestion of the vector with NheI, followed by a Gibson assembly using NEBuilder. Constructs were Sanger sequenced to confirm correct cloning. Lentivirus was made as described above and concentrated 50X. 300,000 SH-SY5Y cells were transduced with 10uL of concentrated lentivirus in media containing 5ug/mL polybrene and seeded in 6-well plates. 2 days after transduction, cells were treated with media containing 15ug/mL blasticidin HCl and selected for at least 3 days until the non-transduced cells were died. Lysate was collected in 1X PLB (Promega) and stored at -80°C prior to performing the luciferase assay. Genomic DNA was also isolated to determine lentiviral integration copy number for luciferase signal normalization. Luciferase assays were performed using a Tecan Infinite M1000 plate reader. Relative luciferase units (RLU) were normalized by both genomic lentiviral copy number and cell lysate. To determine lentiviral copy number, a qPCR was performed using primers that amplify part of the luciferase gene. A standard curve was obtained using a plasmid dilution series. Genomic DNA input was normalized using primers to the intron of WPRE. Cell lysate concentrations were determined using a Pierce Microplate BCA Protein Assay Kit – Reducing Agent Compatible (Thermo Fisher Scientific).

Epigenomic data generation and processing

RNA-seq data generation and primary processing

RNA-seq on the neuronal samples was performed as such. Total RNA was collected using Trizol (Invitrogen) followed by cleanup using RNA Clean and Concentrator (Zymo) using the manufacturer’s protocol. Samples were then QCed using bioanalyzer and subjected to paired end sequencing (BGI platform).

RNA-seq on Astrocyte biological replicates was performed using the Lexogen Quant-seq 3’ mRNA-seq Library Prep Kit FWD for Illumina protocol (cat# 015.96). Briefly, mRNA was isolated and reverse transcribed from 500 ng of total RNA. Double-stranded cDNA was synthesized and i7 adapters for Illumina sequencing were added during PCR amplification.

RNA-seq libraries were sequenced on an Illumina HiSeq 4000 instrument at a depth of 30 million reads per sample.

RNA-seq data from HEK293T cells were processed from the raw read data from Aktas, et al 2017[3], (SRR3997504-7).

Once reads were sequenced or extracted, single end reads were mapped to the hg19 reference genome with GRCh37 Ensembl annotations using STAR aligner (version 2.5.4b)[4] using default parameters. Sample expression counts and transcripts per million (TPM) values were generated using RSEM (version 1.3.0)[5] and default parameters. Conversion between Ensemble IDs and HGNC symbols was performed using the biothings api client (https://biothings.io/) python package v0.2.6. Cell type-specific genes were defined as genes expressed at a TPM>1 across both biological replicates in a single cell type and at a TPM <1 in all other cell types.

RNA-seq differential analysis

Tximport[6] v1.14.0 R package was used to import RSEM counts (see section "RNA-seq data generation and primary processing") into R environment, and R package DeSeq2[7] v1.26.0 was used to call differentially expressed genes. Finally, differential gene TPM values were visualized in heatmaps using the R package pheatmap v.10.12[8]. GO term enrichment for different cell-types was determined via clusterProfileR[9] v.3.14.0. GO[10,11], Reactome[12], and MSigDB[13,14] genesets were utilized in this analysis.

Fast-ATAC sequencing data generation and primary processing

Fast-ATAC sequencing on astrocyte biological replicates was performed as previously described[15]. Briefly, 55,000 viable cells were lysed with digitonin as a detergent and pelleted by centrifugation at 500 g force for 5 minutes at 4C. The nuclei pellet was resuspended in 50 uL of transposase mixture (25 uL 2x TD buffer, 2.5 uL of TDE1, 16.5 uL PBS, 0.5 uL 1% digitonin, 0.5 uL 10% Tween-20, 5 uL nuclease-free water). Transposition reactions were incubated at 37°C for 30 minutes in an Eppendorf ThermoMixer with agitation at 1000 RPM. Transposed DNA was purified using a Zymo DNA Clean and Concentrator-5 Kit (cat# D4014) and purified DNA was eluted in 20 ul elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified, in accordance to published protocols[16] with modified primers[17]. Libraries were quantified using qPCR prior to sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on an Illumina HiSeq 4000 at a depth of 50 million reads per sample.

ATAC-seq read alignment, quality filtering, duplicate removal, transposase shifting, peak calling, and signal generation were all performed through the ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline). Briefly, adapter sequences were trimmed, sequences were mapped to the hg19 reference genome using Bowtie2[18] v2.3.4.1 (-X2000), poor quality reads were removed, PCR duplicates were removed (Picard Tools[19] v2.24.0 MarkDuplicates), chrM reads were removed, and read ends were shifted +4 on the positive strand or -5 on the negative strand to produce a set of filtered high-quality reads. These reads were put through MACS2[20] v2.1.1 to get peak calls and signal files. Finally, IDR analysis was run on the two replicate peak files to produce an IDR peak file that is the reproducible set of peaks across both replicates. The full pipeline can be found on the ENCODE portal.

Differential ATAC peak analysis

ATAC seq peaks were processed for differential expression using a pipeline described here[21] with modifications. R package DeSeq2[7] was used to determine differential counts in ATAC peaks. Briefly, consensus peak regions were established using the R package GenomicRanges (v1.48.0), then the number of ATAC peaks in these peak regions was determined using R package Rsubread (v2.0.0). R package pheatmap was used to plot differential ATAC peaks by cell-type. Additionally, R package ChIPseeker[22] v1.22.0 was used to annotate ATAC peaks were nearest genes for GO term enrichment.

HiChIP data generation and primary processing

The HiChIP protocol was performed for Astrocytes, ESC cells, N-D2, N-D4, N-D10, and N-D28 as previously described[23] using antibody H3K27ac (Abcam, ab4729) at 1ug/ul with the following modifications. Samples were sheared using a Covaris E220 using the following parameters: Fill Level = 10, Duty Cycle = 5, PIP = 140, Cycles/Burst = 200, Time = 4 minutes and then clarified by centrifugation for 15 minutes at 16100 g force at 4° C. H3K27ac antibody was diluted as such: 10X volume of ChIP Dilution Buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM NaCl, water) was added to 4 ug of H3K27ac antibody, and chromatin was incubated overnight. The chromatin-antibody complex was captured with 34 uL Protein A beads (Thermo Fisher). Qubit quantification following ChIP ranged from 125-150 ng. The amount of Tn5 used and number of PCR cycles performed were based on the post-ChIP Qubit amounts, as previously described[23].

HiChIP samples were size selected by PAGE purification (300-700 bp) for effective paired-end tag mapping and where therefore removed of all primer contamination. All libraries were sequenced on the Illumina NovaSeq 6000 instrument to an average read depth of 300 million total reads.

HiChIP paired-end reads were aligned to the hg19 genome using the HiC-Pro pipeline[24] v2.11.1. Default settings were used to remove duplicate reads, assign reads to MboI restriction fragments, filter for valid interactions, and generate binned interaction matrices. HiC-Pro filtered reads were then processed using hichipper[25] v0.7.0 using the {EACH, ALL} settings to call HiChIP peaks in MboI restriction fragments. HiC-Pro valid interaction pairs and hichipper HiChIP peaks were then processed using FitHiChIP[26] v7.0.0 to call significant chromatin contacts using the default settings except for the following: MappSize=500, IntType=3, BINSIZE=5000, QVALUE=0.01, UseP2PBackgrnd=0, Draw=1, TimeProf=1.

HiChIP differential analysis

Differential loop analysis was done using the R package diffloop[27] (v1.10.0). First, a loop object or matrix was created with each row representing a loop (two 5kb DNA segments that are linked together in *cis*- formation) and each column representing a cell replicate. Values are the number of reads counted per loop rows as loops, columns as cell type samples and values as the number of read counts part of the loop. Differential loops are called using limma v3.42.0, similar to how differential RNAseq analysis. Heatmap of results is plotted using R package pheatmap v1.0.12, and the R function "scale" was used to Z-score by column (cell type).

Virtual 4C plot generation

Virtual 4C plots were generated to depict looping relationships extracted from HiChIP centered on a chosen 5kb region, typically containing the TSS of a gene of interest or the SNV of interest for gene and SNV-centric approaches, respectively. First, a bed file is made from the *.abs.bed and *.matrix output files from HiCPro (v.2.11.1) to derive a count matrix of interactions between 5kB bins. Counts were normalized by the number of validate pairs reported by HiC-Pro for each cell type. Interaction frequency is then plotted in R for all samples of interest.

Track plot generation

Tracks are plotted in WashU Epigenome Browser (https://epigenomegateway.wustl.edu/). Tracks were made using hg19 reference genome. Gene marker tracks were extracted from UCSC browser (https://genome.ucsc.edu/). HiChIP loops were displayed as 'longrange' tracks extracted from an overlay of both HiCPro bed files and FitHiChIP bed files. Each purple arc represents one loop. ATAC peaks were displayed in 'bigwig' file format where the peak height on the track corresponds to a normalized read frequency at a given genomic region. MPRA tracks were generated by creating bigwig files containing peaks at the genomic location of the daSNVs, where height of the peak corresponds to the absolute value of the $\log_2$(fold-change) of alternate over reference allele.

General Analysis of epigenomics data

Reference genome hg19 and GENCODE v19[28] were used. Conversions of ENSEMBL[29] ids to gene symbols were doing using python package biothings v0.2.6[30]. For transcription factors and motifs, the HOCOMOCO v11[31] database were used. Heatmaps were made using pheatmap v1.0.12.

Fast-ATAC sequencing data generation and primary processing

Fast-ATAC sequencing on astrocyte biological replicates was performed as previously described[15]. Briefly, 55,000 viable cells were lysed with digitonin as a detergent and pelleted by centrifugation at 500 g force for 5 minutes at 4C. The nuclei pellet was resuspended in 50 uL of transposase mixture (25 uL 2x TD buffer, 2.5 uL of TDE1, 16.5 uL PBS, 0.5 uL 1% digitonin, 0.5 uL 10% Tween-20, 5 uL nuclease-free water). Transposition reactions were incubated at 37°C for 30 minutes in an Eppendorf ThermoMixer with agitation at 1000 RPM. Transposed DNA was purified using a Zymo DNA Clean and Concentrator-5 Kit (cat# D4014) and purified DNA was eluted in 20 ul elution buffer (10 mM Tris-HCl, pH 8). Transposed fragments were amplified and purified, in accordance to published protocols[16] with modified primers[17]. Libraries were quantified using qPCR prior to sequencing. All Fast-ATAC libraries were sequenced using paired-end, dual-index sequencing on an Illumina HiSeq 4000 at a depth of 50 million reads per sample.

ATAC-seq read alignment, quality filtering, duplicate removal, transposase shifting, peak calling, and signal generation were all performed through the ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline). Briefly, adapter sequences were trimmed, sequences were mapped to the hg19 reference genome using Bowtie2[18] v2.3.4.1 (-X2000), poor quality reads were removed, PCR duplicates were removed (Picard Tools[19] v2.24.0 MarkDuplicates), chrM reads were removed, and read ends were shifted +4 on the positive strand or -5 on the negative strand to produce a set of filtered high-quality reads. These reads were put through MACS2[20] v2.1.1 to get peak calls and signal files. Finally, IDR analysis was run on the two replicate peak files to produce an IDR peak file that is the reproducible set of peaks across both replicates. The full pipeline can be found on the ENCODE portal.

MotifBreakR analysis

R package MotifBreakR[32] v.2.10.2 was used to determine the identity of motifs broken or gained by a SNV and magnitude of the allele change. daSNVs were mapped to rsIDs using SNVlocs.Hsapiens.dbSNV142.GRCh37. The HOCOMOCO database was used as a reference for motif PWM (position-weight matrices). A “broken motif” indicates the SNP of interest has a lower match score to the PWM when using the alternate allele versus when using the reference allele. A "gained" motif indicates the opposite. Histograms and density plots were generated using ggplot2. daSNVs were tested for enrichment of broken/gained motifs using a hypergeometric test. Heatmap showing normalized enrichment scores from the hypergeometric test across different neuropsychiatric diseases were shown. Results are shown in **Extended Data Fig. 2F, table S6B.**

Activity-by-Contact model to predict SNV-gene targets

The Activity-by-Contact model[33] (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction v0.2.0) was used as an orthogonal means to predict SNV-gene targets. The process was followed as described[34]. Briefly, candidate regions, or putative enhancer elements were defined by ATAC-seq peaks previously called. Activity, or the number of ATAC-seq reads in these candidate regions, was quantified, and gene body regions were defined via Gencode.v19 annotations. ABC scored were computed by combining activity and contact, defined by FitHiChIP loops, for each cell type. The element-gene prediction pairs were used to assign ABC-predicted target genes to each daSNV. Results are listed for the daSNVs in the column abc_genes in **Data S3**.

Allele specific analysis

Allele specific for HiChIP and ATAC was performed as described in[35] and in accordance with pipelines described by GATK[36] (https://gatk.broadinstitute.org/hc/en-us) v4.1.9.0. Briefly, allele-specific bam files were generated from the initial using the bwa package. Picard[19] v2.24.0 was used to build bam indices, remove duplicates, and sort the resulting bam file. GATK BaseRecalibrator was called to generate a recalibration table based on covariates extracted from known SNV sites (dbSNV_138.hg19, 1000G_phase1.snps.high_confidence.hg19, 1000G_phase1.indels.hg19, and Mills_and_1000_gold_standard.indels.hg19). A collated list of all loop regions and all ATAC peaks were collated for asHiChIP and asATAC analysis, respectively, and used to focus analysis on regions of interest. The base quality score recalibration table was applied to the SNVs using the ApplyBQSR command. HaplotypeCaller was using to generate an allele specific vcf file per sample. Samples were aggregated using the CombineGVCFs command, and genotypes using GentypeVCFs to create the raw overall SNV vcf. Variant recalibration was called with the following parameters:

```
        --resource:hapmap,known=false,training=true,truth=true,prior=15.0
${refSNV_path}/hapmap_3.3.hg19.sites.vcf \
        --resource:omni,known=false,training=true,truth=false,prior=12.0
${refSNV_path}/1000G_omni2.5.hg19.sites.vcf \
        --resource:1000G,known=false,training=true,truth=false,prior=10.0
${refSNV_path}/1000G_phase1.snps.high_confidence.hg19.sites.vcf \
        --resource:dbsnp,known=true,training=false,truth=false,prior=2.0
${refSNV_path}/dbsnp_138.hg19.vcf \
        -an DP -an QD -an FS -an SOR -an MQ -an ReadPosRankSum \
        -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
  -mode SNV
```

Finally, ApplyVQSR was using to generate the final, unfiltered SNV vcf file. vcf files were filtered for SNV locations were a depth count (DP) >=10, alleles with only biallelic SNV sites (GT:0/1), and a minimum reference or alternative allele count (AD) >=2. A binomial test was used to determine if the reference allele count was significantly different from the alternate allele count, and p-values were FDR corrected based on the total number of qualified SNPs based on the DP, GT, and AD filtering. An FDR-corrected binomial p-value threshold of 0.05 was used to determine allele specificity. Hypergeometric tests were performed to determine whether MPRA daSNVs were enriched for asATAC or asHiChIP sites per tissue, using a background of possible MPRA tested SNPs that were called as heterozygous SNP post DP, GT, and AD filtering.

Density plots depicting allele-specific epigenomic signal were generated by counting reads within 150bp up or downstream of SNV of interest for the reference and alternate allele sequences. Counting was done using samtools[37]. Results are listed for the daSNVs in the respective columns in **Data S3**.

<u>Therapeutic analysis</u>

Several drug databases were used to uncover the therapeutic modulation potential of our neuropsychiatric-prioritized genes. First, we curated a list of 166 psychiatric drugs with ATC codes indicated for psychiatric disease. These included 67 antipsychotics, 62 antidepressants, 36 anxiolytic agents, and 1 dopaminergic agent. Additionally, 6798 drugs from The Drug

Repurposing Hub (http://www.broadinstitute.org/repurposing), where extracted, which contained 374 drugs indicated for neuropsychiatric conditions.

Additionally, CMAP[38] (https://www.broadinstitute.org/connectivity-map-cmap) was used to determine the potential for re-purposable drugs to modulate expression of our prioritized genes. As CMAP contains differential expression activity across multiple cell lines in various drug treatment conditions, only HEK293T, neural progenitor cells (NPC), and differentiated neurons (NEU) cells were used in this study. A CMAP level 5 Z-score of 2.5 or -2.5 was used as a cutoff for upregulated and downregulated genes in a drug perturbation condition, respectively. Overall, 295 MPRA-prioritized genes were found to be upregulated by psychiatric drugs, and 173 genes were found to be downregulated by psychiatric drugs. Results are found in **Extended Data Fig. 8**, with **table S11** showing a full list of prioritized drug targets.

Colocalization of GTEx eQTLs and intersection with MPRA results

Colocalization analysis was performed by integrating the disease-associated regulatory risk variants found from MPRA, 114 GWAS and 45 UKBB variant-trait association studies, and 49 GTEx eQTL tissue datasets. Variants were annotated with association summary statistics and filtered by p-value. GWAS p-values were required to be < 5e-8. GWAS studies were only included if they contained at least one daSNVs. Tissue-specific genes were only included if the eQTL colocalized with at least one variant and passed a tissue-specific FDR cutoff. Colocalization was done via enloc[39] (https://github.com/xqwen/integrative) and PhenomeXcan[40] (https://github.com/hakyimlab/phenomexcan) to derive significant GWAS-tissue eQTL colocalizations with at least one MPRA variant genome-wide signicant in both study types. Results are found in **table S8.**

VA cohort analysis of serum magnesium levels in chronic kidney disease

The U.S. Department of Veterans Affairs (VA) healthcare system serves over 9 million veterans at over 1,200 Veterans Health Administration sites of care throughout the United States and U.S. territories[41]. All VA facilities were included in this analysis, and analysis was inspired by prior work[42,43]. Patients were examined who'd had a serum magnesium level measured between January 1, 2021 and December 31, 2021. No patients were excluded. Serum magnesium levels were confirmed in multiple ways to confirm they were a valid test, including using LOINC codes, and ensured that they were measured in the correct unit (mg/dL). All serum magnesium levels that were noted by the lab as partially hemolyzed were removed. Ultimately, n=846,795 patients were in the dataset. If a patient had multiple serum magnesium results, the average was computed and used. Patients' ages and genders were documented and were noted to be predominately male and between ages 45-85. Eight neuropsychiatric conditions were identified in the patients using one year of prior ICD-10 codes that were normalized appropriately. These were Alzheimer's Disease, ADHD, Bipolar Disorder, Generalized Anxiety Disorder, Major Depressive Disorder, Obsessive Compulsive Disorder, Parkinson's Disease and Schizophrenia. Chronic kidney disease was included as a control with well documented relationship with magnesium levels. Relative disease prevalence for serum magnesium levels in the bottom 10th and upper 10th deciles, as well as bottom and upper of six quantiles. Given that Alcohol Use Disorder has a well-known effect on Serum Magnesium levels, patients with this condition were identified for subsequent additional analysis to remove a potential confounder. Significance was determined by linear regression.

All analyses and data visualization were conducted with R and Excel (Microsoft). Results are shown in **Extended Data Fig. 6.**

Lead Contacts

Further information and requests for resources and reagents should be directed to and will be fulfilled by Lead Contact Paul A. Khavari (khavari@stanford.edu).

Extended Figure Legends

Fig. S1. MPRA QC Statistics. (A) Bar chart showing number of reads per MPRA sample in log scale. Replicates are the number following the "R" prefix. Cell type abbreviations are as follows: AST= astrocytes; ES=hESC or human embryonic stem cell; A-NPC = anterior neural progenitor cell; P-NSC = posterior neural progenitor cell; N-DX = induced neuron of day X. Histograms showing barcodes per sequence in the **(B)** plasmid (prior to lentiviral infection) and **(C)** RNA library (extract post infection). **(D)** Power analysis for different levels of barcodes power for at 4 different log2-fold change thresholds (1.2, 1.5, 2, 3), for a total of 20 variants, simulated 5 times

with 5, 10, 20, 50, and 100 barcodes each. **(E)** QQ plots showing the -log10 empirical vs theoretical p-values derived from MPRAnalyze for HEK293T as a cell-type example. The red line is (y=x). **(F)** Histograms showing barcodes per sequence in the RNA library, by cell-type. **(G)** Heatmap showing Pearson count correlation between replicates for all cell types, conditions, and replicates.

Fig. S2. Epigenetics study of the role of transcription regulation in neuropsychiatric diseases. (A) Heatmap showing TF footprints that are enriched in cell types; color scale is normalized count values. **(B)** GO Biological Process dotplot depicting enrichment terms for genes closest to ATAC accessible peaks found across ES-derived neuronal differentiation. The size of the dot is the number of genes in the GO geneset and the color indicates FDR-adjusted p-values. **(C)** Bar chart showing frequency of loop types in promoters and promoter interaction anchor loops (putative enhancers) derived from HiChIP data. Type 1: where an enhancer is linked to a distal gene and the nearest gene, Type 2: where an enhancer is linked only to a distal gene, Type 3: where an enhancer is looped to the closest gene. **(D)** % of P-P (promoter-promoter) and P-PIR (promoter to promoter interaction regions) loops per cell type found via HiChIP. **(E)** Cumulative distribution curves of distance between loop anchors for the different tissues. **(F)** Heatmap (left) showing normalized enrichment scores of motifs broken or gained by SNVs associated with different neuropsychiatric diseases derived from MotifBreakR, relative to a background of other neuropsychiatric diseases. The * refers to motifs that are significantly broken (p-value < 0.10, Fisher's exact test) in daSNVs compared to non-daSNVs for a specific disease. The heatmap (right) shows the log TPM expression values of these transcription factors in different neuronal cell lines and cell lines. **(G)** Scatterplot comparing log-2 fold changes (n=206 variants) for the MPRA dataset (y-axis) with an external Zhang, et al 2020 allele specific open chromatin dataset (x-axis), with a Pearson correlation of 0.48, p-value 1.7x10^{-13}.

Fig. S3. eGene Network Analysis of additional diseases. eGene networks for the additional neuropsychiatric diseases with at least 20 eGenes (from left to right, top to bottom): MDD, BPD, OCD, ADHD, and GAD.

Fig. S4. *POU5F1/OCT4* Vignette. (A) Tracks for the *POU5F1/OCT4* TF gene, where the peak tracks show the logFC change from cell-type specific MPRA for the daSNVs, and the bottom loop track shows the looping data for N-D2 cell type. Boxplots depicting ratios of cDNA to plasmid counts for reference versus alternate allele for SNVS **(B)** rs28428768, **(C)** rs2442722**,** **(D)** rs35735140, and **(E)** rs3134944, where the center line is the median of each MPRA normalized ratio (n=10 genomic instances each); box limits are the upper and lower quartiles, whiskers are the 1.5x interquartile range, and points shown are outliers. Ratios are normalized to the median reference value for each cell type. Significant associations found by MPRAnalyze (FDR < 0.05) are shown with an asterisk*.

Fig. S5. Association between serum magnesium levels and relative psychiatric disease incidence in a VA cohort.

(A) Relative disease prevalence for serum magnesium levels in the bottom 10th and upper 10th deciles. The 10th decile of serum magnesium are values < 1.6 mg/dL and the 90th decile of serum magnesium are values > 2.4 mg/dL. ** indicates significance between the two proportion based on a two-sided 2-proportion z-test FDR-corrected p<0.05 for a given disease. **(B)** Relative prevalence of diseases by serum magnesium levels in the VA cohort. The above graph includes all patients age 45-85, n=846795. The below graph removes all patients who were diagnosed with Alcohol Use Disorder, n=618692. Cohort was partitioned by serum magnesium levels into 6 quantiles and the prevalence of each disease was calculated within the quantile. Relative prevalence is calculated as the prevalence normalized to the disease prevalence in the entire cohort. Significance is determined by linear regression with the null hypothesis beta =0, with p-values < 0.10 showed in solid. Abbreviations of disease are as follows: ADHD=Attention Deficit Hyperactivity Disorder, PD=Panic Disorder, GAD=Generalized Anxiety Disorder, BPD=Bipolar Disorder, MDD=Major Depressive Disorder, OCD=Obsessive Compulsive Disorder, SCZ=Schizophrenia, AD=Alzheimer's Disease, CKD=Chronic Kidney Disease

Fig. S6 *RERE* Vignette. (A) Tracks for gene *RERE*, where the MPRA peak tracks show the logFC change from cell-type specific MPRA for the daSNVs, and the bottom ATAC peak show accessibility profiles for all cell types. Box-and-whiskers plots depicting ratios of cDNA to plasmid counts for reference versus alternate allele for daSNVs **(B)** rs301806, the SNV of interest and **(C)** rs301807, as comparison, where the center line is the median of each MPRA normalized ratio (each point is a genomic instance with at least one count), box limits are the upper and lower quartiles, whiskers are the 1.5x interquartile range, and points shown are outliers. Ratios are normalized to the median reference value for each cell type. Additionally, MotifBreakR results are shown for **(D)** rs301806 (above) and rs301807 (below), depicting loss of RUNX1 motif in rs301806, and no RUNX1 motif present at rs301807 loci. (**E**) ChIP PCR for the transcription factor RUNX1 with n=replicates, * indicated significance of two-sided paired t-test p-value between the reference and alternate allele for the two SNPs.

Fig. S7. CMAP drug perturbation analysis. Drug-eGene networks for **(A)** SCZ, **(B)** BPD, and **(C)** MDD. Linkages between eGene to drug indicate that the drug significantly upregulated (red) or downregulates (blue) the expression of that gene in neuro-relevant cell lines in CMAP. Genes (diamonds) are outlined based on the MPRA log fold change direction (red: positive, blue: negative). Drugs (ellipses) are color coded by drug type. Drug-gene pairs towards the left side of the map indicate the MPRA and expression vectors point in the same direction (putatively side effect causing variants); drug-gene pairs towards the right side of the map indicate MPRA and expression vectors pointing in the opposite direction (putatively therapeutic effects).

Fig. S8. Gene concordance for variant annotation approaches.
(A) Distribution of # daSNVs for a GTEx eGene annotations show eGenes are on average, linked to five daSNVs**. (B)** Density plot showing the distribution of daSNV-to-eGene distance with the mean depicted as a vertical red dotted line at 20kB. **(C)** pie chart showing gene annotation concordance between the different annotation of daSNVs, indicating almost a half of GWAS gene annotations do not match expression or chromatin-based gene linkages. (**D**) Enrichment map made via ClusterProfiler showing GO Molecular Functions enriched in genes linked to daSNVs.

Supplementary Tables

Attached as supplementary_tables.xlsx

Table S1. Data Summary

Summary of the experiments, cells and cell lines used

See supplementary_tables.xlsx

Table S2. Comparison to External Variant Prediction

Scoring files for the 2221 MPRA variants for both DeepSea and gkmSVM prediction.

See supplementary_tables.xlsx

Table S3. Comparison to External ATAC

Comparison of ATAC data to Zhang (PMID: 32732423), Inoue (PMID: 31631012), Song (PMID: 31367015) data.

See supplementary_tables.xlsx

Table S4 Comparison to External Looping Data

Comparison of HiChIP data to Song (PMID: 31367015) pcHiC data via A) anchors and B loops.

See supplementary_tables.xlsx

Table S5. GWAS enrichment odds ratio of daSNVs in cell-type specific ATAC and HiChIP regions

Details of enrichment odds ratio of the daSNVs by disease over differential loop regions that were filtered by ATAC peaks. Type 2 diabetes mellitus (T2DM) was used as a control and indicated no enrichment. Enrichment was concentrated to the neuronal stem cells and the embryonic stem cell neuronal lineages.

See supplementary_tables.xlsx

Table S6. SNP-Motif analysis summary table

(A) Detailed MotifBreakR results listing motifs broken and gained and scores associated with each daSNV, as well as (B) summative analysis stating significant enrichment for motifs gained or broken (pval_g or pval_b, respectively) for each motif in each disease.

See supplementary_tables.xlsx

Table S7. Luciferase assay results

daSNVs (n=10 assayed) were run through luciferase assay in SH-SY5Y cells for 4 replicates per allele per SNP) both in episomal and lentiviral formats. Reference and alternate allele luciferase signal (firefly to Renilla ratio, normalized to empty vector controls) was reported.

See supplementary_tables.xlsx

Table S8. Colocalization analyses of MPRA hits with GTEx

Colocalization results based on annotation of MPRA variants with GTEx and GWAS summary statistics, following by filtering and colocalization steps.
See supplementary_tables.xlsx

Table S9. BrainMap (single cell cortical brain data) Annotation for gene linked to daSNVs

See supplementary_tables.xlsx

Table S10. SCZ disease genes linked to protein coding variants and daSNVs.

List of SCZ-associated genes (n=7) prioritized for protein coding and/or causal variants based on a review of SCZ genetic literature. All genes listed have epigenomic data that links them to SNVs significant in our MPRA study. daSNVs (column 3) are MPRA-significant SNVs that loop to the gene of interest in neural cell types based on HiChIP data. Is eQTL (column 4) is a Boolean indicator of whether or not GTEx, PsychENCODE, and eQTLgen list the daSNVs are an eQTL in brain-relevant tissues (where tissue-specific information is available). SCHEMA's meta analysis p-value, adjusted p-value, protein truncating variants' (PTV) case-control p-value, PTV odds ratio (OR) are shown (columns 5-8). Protein coding mutations (column 9) are missense mutations/PTVs notated within SCHEMA analysis. PMIDs (column 6) are for research articles referencing SCZ GWAS studies, Schizophrenia Exome Sequencing Meta-analysis study (SCHEMA), and various gene-centric papers related to schizophrenia.

Annotations:

note 1: C4A is in the MHC loci and, due to high variability in the region is not included in SCHEMA exome analysis
note 2: XPO7 missense variants/PTVs do not reach high enough allele frequency in cases and/or controls in SCHEMA

gene	gene name	daSNVs	is eQTL	Meta pval	Meta p.adjust	CC pval	OR	Missense Mutations + PTVs	PMIDs
C4A	complement factor 4	10 daSNVs	y	Note 1	Note 1	Note 1	5	Note 1	26814963
CACNA1G	Calcium channel, voltage-dependent, T type, alpha 1G subunit	rs2428682	n	4.57e-7	1.54e-3	3.16e-6	4.25	A2108S, Gly529A, c.6060+2T>C, S818AfsTer21, W925Ter, Q968Ter, Leu1050HfsTer38, W1488Ter, L1685RfsTer27, c.5227-2A>G, c.5925+1G>T	SCHEMA
DAGLA	Diacylglycerol Lipase Alpha	9 daSNVs	n	6.87e-5	4.61e-2	8.99e-5	6.02	L401VfsTer8, c.1213-2A>G, Q451Ter, Y497Ter, c.1514+1G>T, R547Ter, c.2171+1G>A, A843CfsTer158, A1032GfsTer5	SCHEMA
MAGI2	Membrane Associated Guanylate Kinase	rs322004	n	6.41e-5	4.47e-2	3.11e-4	8.03	F163LfsTer11, Q1305RfsTer169, R1084Ter, P908LfsTer32, c.2311+2C>A, E531Ter, I473SfsTer4, c.1225+1G>A, E238Ter, W11Ter	SCHEMA
STAG1	Cohesin subunit SA-1	rs900947	n	5.25e-5	4.34e-2	7.56e-5	8.03	E1235Ter, R1206Ter, L927YfsTer15, I839MfsTer56, R131Ter, R51Ter	SCHEMA
SV2A	synaptic vesicle glycoprotein	rs72708145	y	8.21e-5	4.67e-2	8.8e-4	4.42	R390Ter, F718CfsTer17, R507Ter, V486SfsTer13, Q476Ter, E138GfsTer15, R67Ter, R43Ter	31937764, SCHEMA
XPO7	Exportin 7	rs746011, rs11136093, rs11780207	n	7.18e-9	4.34e-5	2e-8	28.1	Note 2	SCHEMA

Table S11. Psychiatric genes druggability prioritization table

This is a prioritization of potential drug targets (8 high, 12 medium, 33 low priority of the 641 possible genes), collated with a combination of various sources of evidence (see README for more information).

See supplementary_tables.xlsx

Table S12. 58 CNS-relevant monogenic diseases and their genes linked by OMIM

See supplementary_tables.xlsx, only referenced in methods

Table S13. RNA-seq TPM values for all tissues

RNA-seq data processed in house and from external sources. HEK293s TPM values were processed from SRR3997504, SRR3997505, SRR3997506, SRR3997507 (Aktas, et al 2017).

See supplementary_tables.xlsx, only referenced in methods

Table S14. 806 Psychiatric codes in the UK Biobank for which GWAS summary statistics were extracted

See supplementary_tables.xlsx, only referenced in methods

Table S15. Primers

See supplementary_tables.xlsx, only referenced in methods

Supplementary Data Descriptions

Data S1. (separate file) LDSC Analysis for Hereditability

Heatmaps and summary statistics for LDSC analysis of hereditability for RNA and ATAC seq data for GWAS summary statistics as described in methods.

Data S2. (separate file) Literature-derived da-SNV Gene Annotations

Literature summary of the 641 da-SNV associated genes, along with PMID and functional annotations.

Data S3. (separate file) daSNV Summary statistics and annotations table

Expanded version of Table 1 all annotations present for all daSNVs. ReadME page includes the list of annotations included for each daSNV including MPRA summary statistics, druggability information, motif changes, allele-specific ATAC or HiChIP, ABC predictions, and UK Biobank phenotypes linked to each daSNV. Also includes annotations for cell types/samples, oligo sequences for the library and controls used.

Data S4. (separate file) Networks of shared putative pathomechanisms in neuropsychiatric disorders.

Full versions of the daSNV (diamond) -gene (ellipses) networks of shared pathomechanisms in neuropsychiatric disease. Genes are color coded by disease of origin, where the green circles represent implicated genes shared between multiple diseases. Genes are linked via StringDB. Networks included are: regulation of cytokine production (GO biological process), sleep issues (from UK Biobank), anhedonia (UK Biobank), and irritability (UK Biobank).

Data S5. (separate file) MPRA cell-condition-specific summary statistics

Summary p-value and log-fold changes for all MPRA tested SNVs calculated using MPRAnalyze.

Bibliography

1. Kovalevich, J. & Langford, D. Considerations for the Use of SH-SY5Y Neroblastoma Cells in Neurobiology. *Methods Mol. Biol.* **1078**, 9–21 (2013).
2. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
3. Aktaş, T. *et al.* DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115–119 (2017).
4. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
5. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 1–16 (2011).
6. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **4**, (2016).
7. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
8. https://CRAN.R-project.org/package=pheatmap.
9. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).
10. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* vol. 25 25–29 (2000).
11. Carbon, S. *et al.* The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
12. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
13. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
14. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
15. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
16. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–9 (2015).
17. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–90 (2015).
18. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
19. Broad Institute. Picard Tools.
20. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 1–9 (2008).
21. https://rockefelleruniversity.github.io/RU_ATAC_Workshop.html.
22. Yu, G., Wang, L. G. & He, Q. Y. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
23. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

24. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, (2015).
25. Lareau, C. A. & Aryee, M. J. Hichipper: A preprocessing pipeline for calling DNA loops from HiChIP data. *Nature Methods* vol. 15 155–156 (2018).
26. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. FitHiChIP: Identification of significant chromatin contacts from HiChIP data. (2018) doi:10.1101/412833.
27. Lareau, C. A. & Aryee, M. J. Diffloop: A computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics* **34**, 672–674 (2018).
28. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
29. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
30. Xin, J. *et al.* High-performance web services for querying gene and variant annotation. *Genome Biol.* **17**, 1–7 (2016).
31. Kulakovskiy, I. V. *et al.* HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
32. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. MotifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
33. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* vol. 51 1664–1669 (2019).
34. https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction.git.
35. Zhang, S. *et al. Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants Downloaded from*. http://science.sciencemag.org/ (2020).
36. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
39. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genet.* **13**, e1006646 (2017).
40. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083 (2020).
41. Hussey, P. S. *et al.* Resources and Capabilities of the Department of Veterans Affairs to Provide Timely and Accessible Care to Veterans. *Rand Heal. Q.* **5**, (2016).
42. Bayat, V. *et al.* Reduced Mortality With Ondansetron Use in SARS-CoV-2-Infected Inpatients. *Open forum Infect. Dis.* **8**, (2021).
43. Holodniy, M. *et al.* Evaluation of Praedico^{TM}, A Next Generation Big DataBiosurveillance Application. *Online J. Public Health Inform.* **7**, (2015).