scientific reports

OPEN

Characterization of enhancer activity in early human neurodevelopment using Massively Parallel Reporter Assay (MPRA) and forebrain organoids

Davide Capauto[1,9], Yifan Wang[2,9], Feinan Wu[1], Scott Norton[1], Jessica Mariani[1], Fumitaka Inoue[3], Gregory E. Crawford[4], Nadav Ahituv[5,6], Alexej Abyzov[2]✉ & Flora M. Vaccarino[1,7,8]✉

Regulation of gene expression through enhancers is one of the major processes shaping the structure and function of the human brain during development. High-throughput assays have predicted thousands of enhancers involved in neurodevelopment, and confirming their activity through orthogonal functional assays is crucial. Here, we utilized Massively Parallel Reporter Assays (MPRAs) in stem cells and forebrain organoids to evaluate the activity of ~ 7000 gene-linked enhancers previously identified in human fetal tissues and brain organoids. We used a Gaussian mixture model to evaluate the contribution of background noise in the measured activity signal to confirm the activity of ~ 35% of the tested enhancers, with most showing temporal-specific activity, suggesting their evolving role in neurodevelopment. The temporal specificity was further supported by the correlation of activity with gene expression. Our findings provide a valuable gene regulatory resource to the scientific community.

It has been over 40 years since the discovery of the first DNA sequence capable of enhancing the transcription of a reporter gene[1]. Since then, many *cis*-regulatory DNA elements known as enhancers have been identified, and their biochemical and functional properties have been extensively investigated. A central role of enhancers is to regulate gene expression by binding transcription factors (TFs) and other regulators able to modulate the transcription of target genes[2]. Enhancers can act independently of the distance and orientation to their target genes[3] and they have specific chromatin features that aid in their genome-wide identification through high-throughput methods. Common techniques used for enhancer identification are DNase I hypersensitivity sequencing (DNase-seq)[4] and the Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq)[5] which exploit the fact that enhancers are free from nucleosomes and are more sensitive to enzymatic treatment. One powerful tool for identifying active enhancers is Chromatin immunoprecipitation-sequencing (ChIP-seq), which utilizes antibodies that recognize specific histone modifications, such as H3K27Ac and H3K4me1, in enhancers flanking nucleosomes[3,6,7]. Additionally, active enhancers were found to be actively transcribed into enhancer RNAs (eRNAs)[8,9], and to form loops with target promoters to exert their regulatory effects on gene expression. In recent years, several high-throughput 3D chromatin conformation techniques, such as Hi-C[10,11], ChiA-PET[12,13], and micro-C[14], have been used to identify physical interactions between potential enhancers and promoters, elucidating the spatial organization of the genome.

[1]Child Study Center, Yale University, New Haven, CT 06520, USA. [2]Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA. [3]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan. [4]Department of Pediatrics, Duke University, Durham, NC 27708, USA. [5]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [6]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [7]Department of Neuroscience, Yale University, New Haven, CT 06520, USA. [8]Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA. [9]These authors contributed equally: Davide Capauto and Yifan Wang. ✉email: abyzov.alexej@mayo.edu; flora.vaccarino@yale.edu

While these approaches are highly informative and allow identifying the location of potential enhancers, none of these represent the definitive proof of their activity. To address this, reporter assays have been widely used to confirm the activity of predicted enhancers by testing their ability to drive the expression of a heterologous reporter gene. Previously, this method was limited to testing one sequence at a time using transgenic technologies[15,16], but the development of high-throughput massively parallel reporter assays such as STARR-seq[17] and MPRA[18–21] has revolutionized the field, enabling the simultaneous testing of thousands of DNA sequences for enhancer activity in a single experiment. More in detail, in the MPRA, candidate regulatory sequences, including promoters or enhancers, are introduced in a construct upstream of a minimal promoter, a gene reporter, and a unique barcode. Once the MPRA construct is introduced into cells, the tested sequences drive the expression of the associated unique barcode. The ratio between the number of transcribed barcodes in the RNA fraction (RNA counts) and the number of barcodes in the DNA fraction (DNA counts) provides a quantitative measurement of the regulatory activity of the tested sequences.

Recently, MPRA experiments have been employed to understand transcriptional patterns driving neuronal differentiation through perturbation of transcription factor binding motifs within regulatory elements[22,23] and for studying the effect of thousands of human specific substitutions on the functional activity of enhancer regions implicated in human brain evolution[24–26], or neuronal-associated diseases[27–29].

Here, we used a genomic-integrated MPRA, the Lentiviral-based Massively Parallel Reporter Assay—LentiMPRA[30]—to evaluate the activity of approximately 7000 putative gene-linked enhancers. These enhancers were previously identified in human induced pluripotent stem cells (iPSC)-derived forebrain organoids and human fetal cortex combining ChIP-seq and Hi-C data[31] and are potentially involved in early human neurodevelopment.

Results

Enhancer selection and lentiMPRA experimental design

To characterize the activity of a set of enhancers putatively involved in early human neurodevelopment, we used lentiMPRA in forebrain organoids to test the activity of a subset of the 96,375 gene-linked enhancers identified by ChIP-seq in our previous study conducted in human fetal cortex and human forebrain organoids[31]. In the original dataset, enhancers were discovered by ChIP-seq, annotated by ChromHMM and linked to putative target genes by proximity and/or external DNA conformation (Hi-C) dataset from stem cells and fetal brains[31]. We generated an MPRA library selecting the most active (e.g., with highest H3K27Ac peak signal) 6989 enhancers in organoids as measured by ChIP-seq (Fig. 1A). Additionally, we included a total of 122 positive control sequences from three different datasets: (i) 87 enhancers from human embryonic stem cells (hESCs) validated using the ChIP-STARR-seq approach[32]; (ii) 21 MPRA-validated enhancers from hESC-derived neuronal progenitor cells (NPCs)[23]; (iii) 35 human brain enhancers from the Vista Enhancer Browser validated by mouse transgenesis[33]. We also included 150 negative controls generated by shuffling the nucleotides of 150 randomly selected candidate enhancer regions. To address the length limitation of oligonucleotide synthesis, we selected a minimal enhancer region of 270 bp by intersecting our enhancers with synonymous information from other datasets, such as p300 ChIP-seq peaks from neuronal cell lines and fetal brain[34,35], DNA hypersensitivity peaks[36] and CAGE analysis from fetal brain and neuronal cell types[35,37–39]. The goal was to identify regions with the highest overlap across these datasets which likely correspond to the core active enhancer regions. If the resulting core regions still exceeded the length of 270 bp, we used FIMO[40] to refine them by identifying a subregion with the highest number of known transcription factor binding sites (TFBSs) (see Supplementary Fig. S1 and Methods). In total, the library included 7261 sequences synthetized along with 15 bp adapters on either side. The lentiMPRA library was amplified, and a minimal promoter and a 15 bp random barcode were placed downstream of each synthesized sequence and cloned into a lentiMPRA vector upstream of the GFP coding sequence (Fig. 1A).

To investigate whether our candidate enhancers can elicit a time-dependent transcriptional response during early neurodevelopment, we infected three induced pluripotent stem cell (iPSC) lines with our lentiMPRA library and measured enhancer activity at the iPSC stage and at two terminal differentiation (TD) time points during forebrain organoid differentiation: an earlier stage (TD0) when organoids were almost exclusively composed of proliferating progenitors and a more mature stage (TD30) when organoids were still harboring progenitors and actively generating layer 5–6 cortical postmitotic neurons (Supplementary Fig. S2 and S3; Methods).

LentiMPRA identifies active enhancers during forebrain organoids differentiation

Using DNA sequencing, we discovered that 95.1% (6907/7261) of the tested enhancer sequences in the original library were successfully recovered, with an average of 39.9 unique barcodes associated with each enhancer sequence (Supplementary Fig. S4). Out of the total 7261 sequences, 94.9% passed stringent quality control (Methods). In the MPRA experiment, enhancer activity is measured as the ratio of transcribed barcode reads (obtained by RNA-seq) to integrated genomic barcode reads (evaluated by DNA-seq). To identify active enhancers in the tested set, we first defined the background distribution for the ratio of RNA/DNA barcodes. We reasoned that the presence of a multitude of TFs in each cell, combined with the non-deterministic (i.e., probabilistic) nature of TF binding motifs and the largely unstudied effects of TF cooperativity, could result in enhancer activity even in randomly shuffled negative control sequences. Consequently, the distribution of activity for the negative control sequences was approximated as a sum of two gaussian distributions, one representing the true background signal and one representing an actual signal from potentially active sequences (Fig. 1B and Supplementary Fig. S5; Methods). Such a bimodal approximation precisely described the observed experimental distribution with an average signal from likely active sequences being roughly 55% stronger than the average background (Fig. 1B). Such an approximation also described the signal distribution for positive controls, suggesting that some of the positive controls are not active, albeit enriched for active enhancers as compared to negative controls (Supplementary Fig. S5).

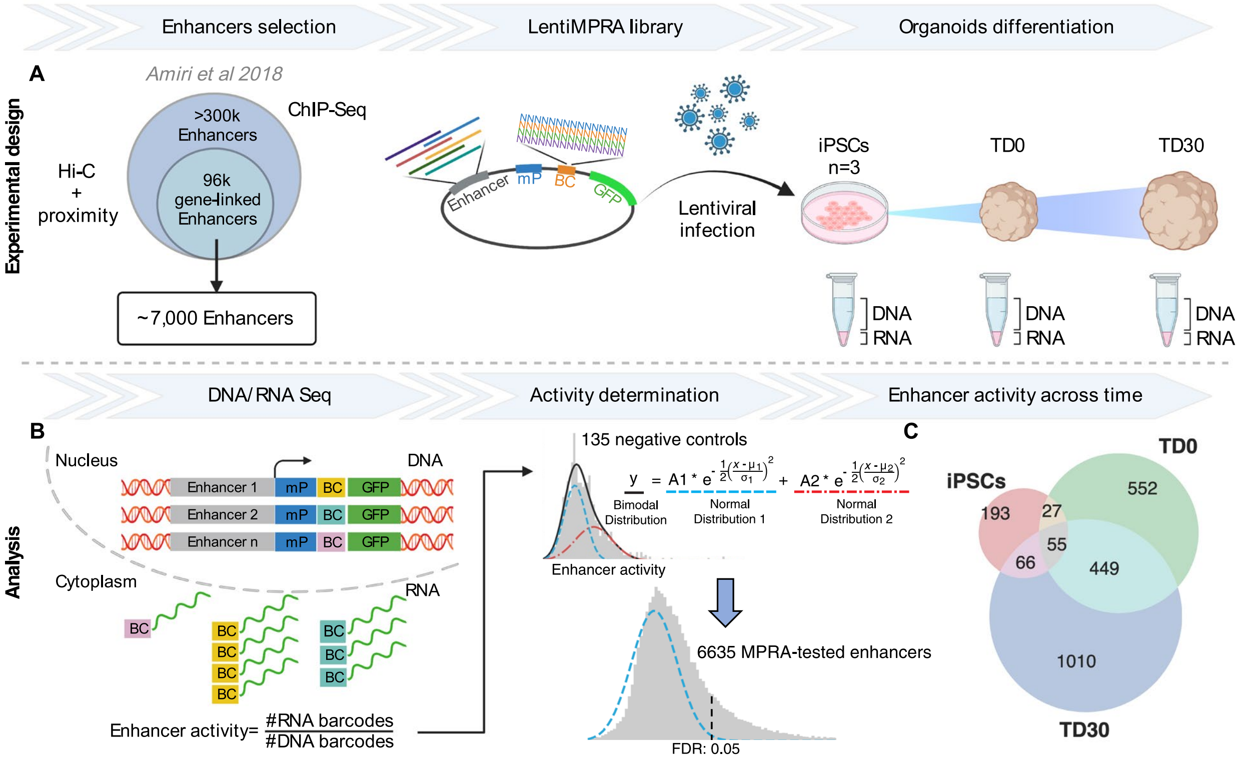

Figure 1. Experimental design and overall lentiMPRA results. (**A**) ~ 7000 candidate enhancer regions were selected among the 96,000 gene-linked enhancers identified by Amiri et al.[31]. The lentiMPRA library was synthesized on a custom array and cloned into a lentiMPRA vector, packaged into lentivirus, and introduced into 3 iPSC lines before organoid differentiation. (**B**) At 3 different time points, DNA/RNA sequencing was used for estimating the enhancer activity from the ratio of corresponding barcodes. Enhancers were considered active by MPRA if their activity was significantly above the background (blue curve) derived from a gaussian mixture model of the activity of negative controls. (**C**) The Venn diagram shows counts of MPRA-active enhancers across time points. BC, barcode; LTR, long terminal repeat; mP, minimal promoter; TD, terminal differentiation.

MPRA-active enhancers were defined as those having a signal significantly above the background distribution (Fig. 1B). Altogether, 34.8% of the enhancers were active in at least one time point (Supplementary Table S1). Of the core validated enhancers, 1755 (74.6%) exhibited activity at a specific time point, reflecting a stage-specific epigenomic regulation: 193 were active only in iPSCs, 552 at TD0, and 1010 at TD30. More than 25% of active enhancers were shared between time points, with a large proportion being active at both TD0 and TD30 (449), indicating that some regulatory elements may play a role in both early and late stages of neurodevelopment (Fig. 1C**).**

The background and the actual signal distribution were wide and overlapped significantly, limiting a clear discrimination between active and inactive enhancers, and reducing the power of the MPRA approach (Fig. 1B). Based on this overlap and a selected *p*-value threshold, we estimated that we were only powered to validate about 35% of truly active enhancers. To elaborate on the missing fraction of true enhancers, we clustered all tested enhancers by their activity profile (RNA/DNA ratio) across samples in the MPRA experiment. This clustering approach revealed that almost all enhancers fell into two large clusters, cluster 1 and 2 (Fig. 2A). Cluster 1 had just a few MPRA-active enhancers, while the cluster 2 encompassed almost all MPRA-active enhancers. We interpreted the data as suggesting that the inactive enhancers in cluster 2 were likely validated, but they did not formally pass the significance test due to the low sensitivity of the MPRA assay. This result would also explain the limited reproducibility of MPRA-active enhancers across samples at TD0 and TD30 (Supplementary Fig. S6).

We next intersected MPRA-tested enhancers with external ATAC-seq-derived peaks obtained from bulk and single cell data from fetal brains and forebrain-directed organoids[41,42] as well as from whole-brain organoids[43]. A large proportion (79.1%) of the tested enhancers were present in these external datasets. Consistent with our expectations, we observed that MPRA-active enhancers were enriched in fetal brains and forebrain-directed organoid datasets from Trevino et al. and Ziffra et al. (Fig. 2B inset, Supplementary Fig. S7 and Supplementary Table S2).

To further qualify the nature of MPRA-active enhancers, we compared the numbers of TFBSs within active and inactive enhancers. Our analysis revealed that while positive control enhancers had significantly more TFBSs compared to both scrambled negative controls and tested enhancers (Fig. 2C), we did not observe any significant difference in the number of TFBSs or in the expression of cognate TFs between MPRA-active and -inactive tested enhancers.

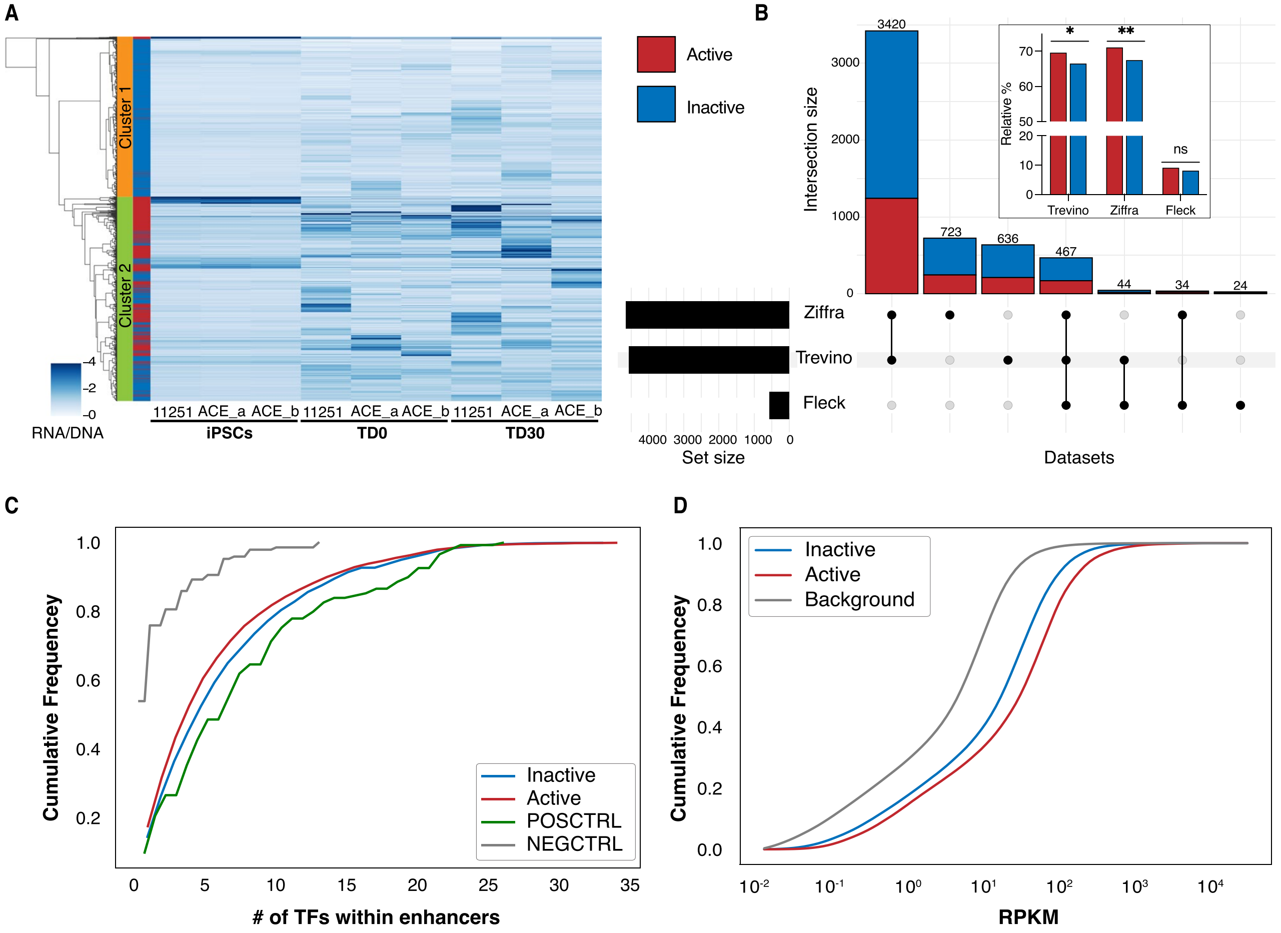

Figure 2. Characteristics of active enhancers. (**A**) Heatmap of enhancer activity (RNA/DNA) of all candidate enhancers tested in MPRA experiment. Clustering was performed using the Ward variance minimization algorithm. The right bar annotates MPRA-active (red) or inactive (blue) enhancers. (**B**) Upset plot showing the number of active (red) and inactive (blue) enhancers overlapping external datasets. In the box, bars represent the fraction of overlapping active and inactive enhancers relative to the total number of active and inactive enhancers. *p*-values were calculated using Fisher's exact test (**p*-value < 0.05; ***p*-value < 0.01). (**C**) Cumulative plot of number of transcription factor binding sites identified by FIMO in positive controls, negative controls, and tested enhancers. No statistically significant difference was observed between active and inactive enhancers. (**D**) Cumulative plot of the expression (RPKM) of genes predicted to be regulated by tested enhancers in TD0 and TD30 organoids in the endogenous genomic context by Amiri et al.[31]. Genes regulated by MPRA-active enhancers have significantly higher expression than inactive enhancers. The background curve represents the expression of all ~ 96,375 gene-linked enhancers from Amiri et al.[31].

We next took advantage of the fact that in the original dataset of Amiri et al. we identified putative target genes for the tested enhancers in their native DNA location, by either proximity in linear DNA and/or 3D DNA conformation datasets[31]. We then compared the MPRA-active and -inactive enhancers with regard to the expression of their linked genes. To increase the statistical power of our analysis, we incorporated bulk RNA-seq data from 88 additional samples from a parallel experiment using genetically distinct iPSC lines (collected at TD0 and TD30, using the same forebrain differentiation protocol) in addition to the samples used in the MPRA experiment. Our analysis revealed that the genes linked to MPRA-active enhancers exhibited significantly higher expression levels than those associated with inactive enhancers (Fig. 2D)[31]. Genes linked to enhancers in both categories had higher expression than "background", i.e., all gene-linked enhancers from Amiri et al.[31]. This likely reflects the selection of the most active enhancers for the MPRA experiment. The fact that MPRA-inactive enhancers had higher target gene expression than the "background" expression was consistent with the low sensitivity of MPRA experiment estimated above, implying that a significant fraction of MPRA-inactive enhancers could actually be active in a different experimental setting.

We then correlated the difference in enhancer activity (measured by lentiMPRA) across time points with the difference in expression of the gene(s) linked to the enhancer in the endogenous genomic context. Comparing enhancer activity and gene expression at TD0 versus the iPSC stage, we found that MPRA-active enhancers upregulated or downregulated at TD0 or TD30 versus the iPSC stage were typically correlated (positively or negatively) with the difference in expression of their endogenous linked genes (Fig. 3A, B and Supplementary Fig. S8, S9). In contrast, MPRA-inactive enhancers rarely showed a correlation with the expression of their

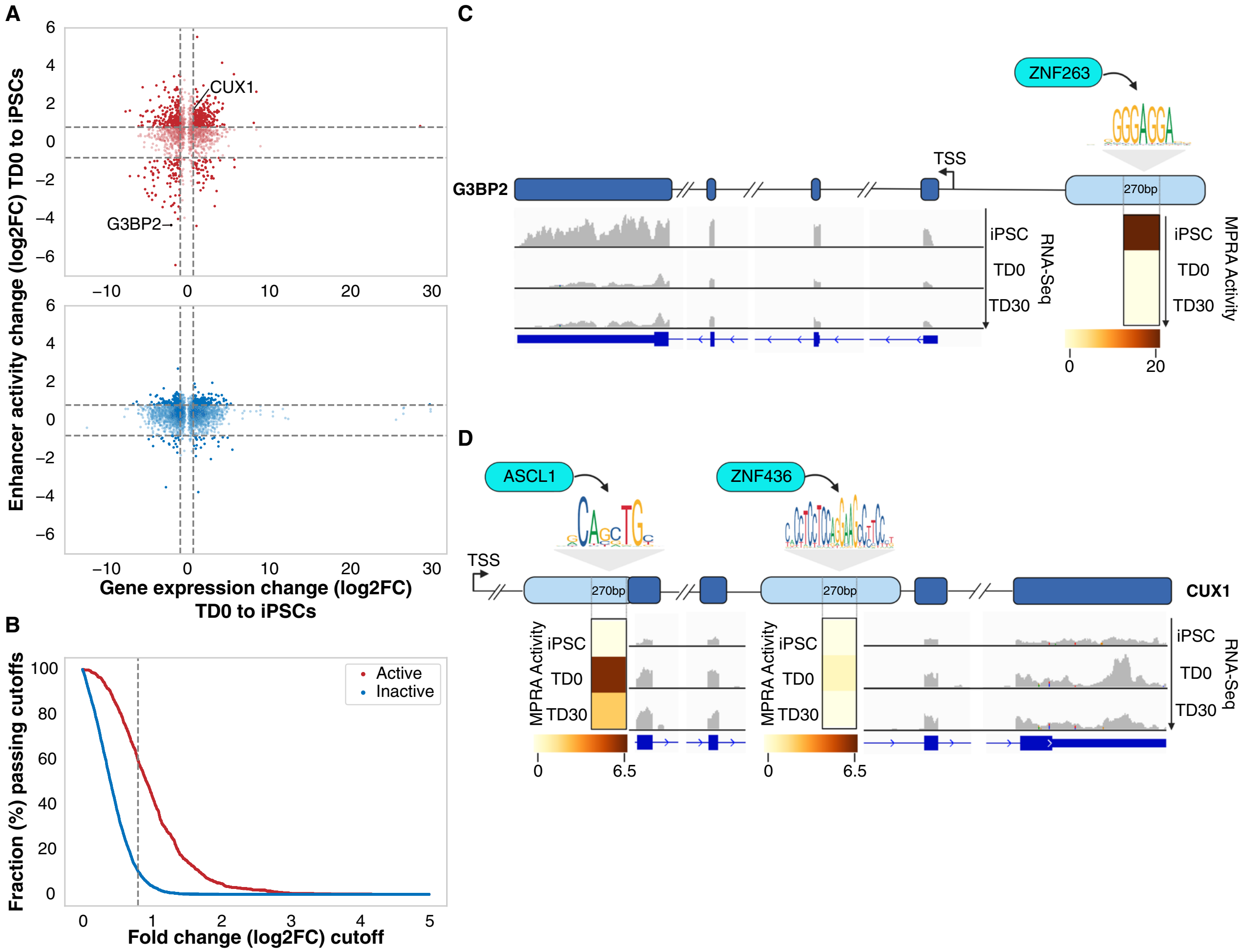

Figure 3. MPRA active enhancers correlate with differences in expression of linked genes across different time points. (**A**) Scatter plots of change in MPRA activity for an MPRA-active enhancer (y-axis) and change of expression for the gene(s) linked to the enhancer in the endogenous genomic context. Activity and expression are compared for TD0 and iPSC. Each dot represents a gene-enhancer pair. iPSC or TD0 active enhancers are shown as red circles in the upper panel; inactive enhancers are represented in the bottom panel as blue circles. (**B**) More correlated enhancer-gene pairs are observed in MPRA-active enhancers compared to inactive enhancers. The x-axis represents the cutoff for the change in enhancer activity and gene expression. The y-axis represents the fraction of enhancer-gene pairs passing cutoff for the change in the enhancer activity and gene expression. The dashed line represents cutoff at log2FC = 0.8, which is the same as marked in panel (**A**). (**C**) Example of an iPSC-only active enhancer (MPRA activity shown in the heatmap) and the expression of the linked gene (G3BP2), with ZNF263 as the predicted binding transcription factor. (**D**) Example of a TD0 active enhancer (MPRA activity shown in the heatmap) and expression of the linked gene (CUX1) with ASCL1, a TF predicted to bind to this enhancer. For comparison, a mildly active enhancer at TD0 for this gene is identified downstream of the active enhancer. TSS, Transcription Start Site.

corresponding genes when comparing either TD0 (Fig. 3A and B) or TD30 with iPSCs (Supplementary Figs. S9 and S10). These observations demonstrate that the MPRA-active enhancers, compared to the inactive ones, were those that, in their appropriate genomic context, were linked to highly expressed genes and that differential MPRA activity predicted differential gene expression during differentiation. Given the nature of the MPRA assay, it can be inferred that the activity of these MPRA enhancers is less dependent on the genomic context. Among the enhancers exclusively active in iPSCs, we found that the most active one (Fig. 3A, Supplementary Table S3) is located upstream of the GTPase-activating protein (SH3 domain)-binding protein 2 (G3BP2) gene, which encodes an RNA binding protein involved in maintaining pluripotency by regulating the transcription factors Oct-4 and Nanog[44] (Fig. 3C). The G3BP2 gene was also upregulated in iPSCs compared to the TD30 stage (Supplementary Fig. S9), consistent with the TD0 MPRA results and with its biological role in maintaining pluripotency. The 270 bp tested enhancer region is potentially bound by seven TFs (Supplementary Table S3), including ZNF263, a TF expressed in pluripotent stem cells[45] and computationally predicted to have the highest affinity for the binding motif (Fig. 3C).

Within the enhancers displaying the highest log2FC activity at TD0 (Fig. 3A), several are linked to genes particularly noteworthy in the context of brain development, suggesting an important role of these enhancers in neuronal functions. The enhancer with the highest differential activity at TD0 compared to iPSC is linked to the Regulator of G Protein Signaling 4 (RGS4) gene. This gene is expressed in the developing nervous system and adult brain, although its function remains unclear[46]. Other examples include ANKRD11, a chromatin regulator modulating gene expression in neural precursor cells[47], and the receptor tyrosine kinase ErbB4[48], which regulates various neuronal processes such as migration, proliferation, and differentiation. Other genes of interest are FAM107A, contributing to neural cell survival, migration, and spine formation[49], and P2RX4, a purinoceptor for ATP involved in key CNS functions[50].

Among the top 50 differentially active enhancers at TD0 versus iPSC, the one displaying the greatest enhancer activity across all the TD0 samples (Supplementary Table S3) partially overlapped with an exon located ~ 400 kb downstream of the transcription start Site (TSS) of the Cut Homeobox 1 (CUX1) gene, a TF playing a critical role in upper layer cortical neuron differentiation, dendrite branching, and synapse formation[51,52]. Despite its role in late differentiation, other works indicated an increase in CUX1 mRNA levels in early progenitors, including radial glia. This suggests that CUX1 could potentially serve as a determinant, in early progenitors, of the identity of the upper cortical layers[53,54]. CUX1-linked enhancer differential activity was also confirmed by comparing TD30 vs iPSC (Supplementary Fig. S9 and Supplementary Table S3). Among the four TFs predicted to bind this active enhancer (Supplementary Table S3) there is ASCL1, which has been previously identified to promote CUX1 expression[55]. Additionally, a second intronic enhancer linked to CUX1, bound by ZNF436, showed mild MPRA activity only at TD0 (Fig. 3D).

Discussion

Spatial and temporal regulation of gene expression through regulatory regions, specifically enhancers, is essential for shaping the structure and function of the human brain during development. A number of biochemical techniques, such as ChIP-seq, ATAC-seq and DNase-seq have been used for identifying enhancers and characterizing their activity in neurodevelopment. Among those, the MPRAs have been adopted to test an enhancer's ability to activate a synthetic reporter in a high throughput and context-independent manner, providing an orthogonal readout of enhancer activity. Most of the MPRA-based studies have primarily concentrated on investigating the functionality of potential regulatory sequences in 2D iPSCs-derived neuronal cells at very early stages of differentiation[22,23] or in human-derived neuronal stem cells[24–26]. In this study, we used lentiMPRA in forebrain organoids to evaluate, over the course of brain development, the activity of ~ 7000 gene-linked enhancers previously identified in human fetal tissues and brain organoids from a combination of histone-mark ChIP-seq and DNA conformational studies. Besides validating 2352 enhancers, our analysis of the MPRA-active enhancers in relation to upstream binding TFs and downstream targets suggested important implications for the developmental biology of the human brain.

By analyzing RNA-seq data from forebrain organoids collected at the same time points, we found that while there was no relation with number of potential TFBSs or TFs expression levels, MPRA-active enhancers are associated with highly expressed genes in the endogenous genomic context. Furthermore, differential enhancer activity, as determined by the lentiMPRA assay, tended to be correlated, in a positive or negative fashion, with differential gene expression across time points of neural differentiation. This provides an interesting insight into the capability of MPRA to detect enhancers associated with genes dynamically regulated during different stages of neurodevelopment.

On a technical side, our study highlights a few essential limitations of MPRA techniques. Perhaps the major limitation of MPRA is its low sensitivity (poor discrimination of signal from noise), estimated to be about 35% in our experiments, which likely explains limited reproducibility across replicates. For the later differentiation time points this could also be due to lentivirus regulatory element transgene silencing, which was previously reported to be an issue over differentiation of stem cells into other cell types[56]. More data are required to precisely assess the variability of this assay across technical replicates and genetically different iPSC lines. In addition, there may be some false negative results in the MPRA assay, since the assay tests a shorter version of the original enhancer, which may not include all the necessary elements required for transcription initiation. For example, MPRA-tested enhancers may lack crucial co-activators that are provided by DNA loop conformation, or may be present in the flanking sequences of the tested region. This may explain why, while detecting a significant enrichment of MPRA-active enhancers with external ATAC-seq datasets obtained in fetal brain or forebrain organoids, a considerable percentage of MPRA-inactive enhancers also overlapped with those external datasets.

Similarly to any bulk assay, using MPRA in a heterogenous system such as organoids, which are characterized by cellular diversity, reveals another caveat of the technique. Given the absence of cell-type information coupled to enhancer activity, it is likely that MPRA may be strongly biased towards revealing those enhancers present in the most abundant cell types. Indeed, when intersecting with scATAC-seq data obtained from the same organoids preparations at TD0 and TD30, there was a significant enrichment of MPRA-active enhancers only in the most abundant cell population, radial glial cells (Supplementary Table S4)[57]. Future studies using single-cell MPRA-seq approaches[58,59] in combination with existing single-cell biochemical assays such as scATAC-seq, could open new avenues to gain a comprehensive understanding of the fine regulatory dynamics occurring within each cell type in complex neuronal differentiation systems.

Materials and methods

Human subjects

Human subjects were recruited through several research projects at the Yale Child Study Center. Written informed consent was obtained from each participant enrolled in the study, and all research was approved by the Yale University Institutional Review Board (HIC# 1104008337) and Yale Center for Clinical Investigation at Yale University and was performed in accordance with the Declaration of Helsinki. Human participants' names and other HIPAA identifiers were removed from all sections of the manuscript, including supplementary information. The participants agreed to data sharing of genomic unidentified data using controlled data access.

All methods were performed according to relevant regulations and guidelines of Yale University, including the Biological Safety Committee and ESCRO committee of Yale University.

LentiMPRA library design

Candidate enhancers were identified in Amiri et al. 2018[31] by chromatin segmentation analyses using H3K27Ac, H3K27me3 and H3K4me3 ChIP-seq datasets obtained in cortical organoids and cerebral cortical tissue from postmortem fetal human brains. Enhancers were linked to putative target genes by proximity and/or by association with promoters using fetal brain 3D chromatin conformation data. From an initial dataset of > 300,000 enhancers, 96,375 were found to be associated with genes and termed gene-linked enhancers (GLE). From the GLE dataset, approximately 7000 enhancers were selected for MPRA based on top activity, as determined by H3K27Ac peak signal. As ChIP-seq enhancers can encompass hundreds to thousands of bases, for the purpose of oligonucleotide synthesis for the MPRA assay, a core of 270 bp region was identified. To achieve this, our chosen enhancers were intersected with enhancers from other datasets to select the region with the highest number of overlaps, and therefore potentially more active. In detail, we used: (i) p300 ChIP-seq peaks from human neuronal cell lines[39] and human fetal cortex[60], (ii) CAGE analysis from brain tissues[37], and (iii) DNase hypersensitivity peaks from neuronal progenitor cells and brain tissue[36]. Finally, if there was no overlap or if the overlapping enhancer region was still too large, we used FIMO[40] to further reduce the size to 270 bp by selecting the region with the highest number of TFBSs, as TF binding is a reliable predictor of enhancer activity. The list of tested enhancers is outlined in Supplementary Table S1.

LentiMPRA library generation

The lentiMPRA plasmid library was constructed as previously described in Gordon et al., 2020[30]. Briefly, the oligonucleotide pool of the 7261 enhancers was synthetized by Twist Bioscience and amplified via two rounds of PCR, first to add the minimal promoter and then to add the barcode, using two sets of adaptors primers, 5BC-AG-f01/r01 and 5BC-AG-f02/r02 respectively (Supplementary Table S5). The amplified fragments were cloned via Gibson assembly (using NEBuilder HiFi DNA Assembly Master Mix; New England BioLabs, cat. No. E2621L) into the SbfI/AgeI site of the pLS-SceI vector (Addgene, cat. No. 137725, a gift from Ahituv's lab) to construct the library. The resulting library was digested with I-SceI (New England BioLabs, cat. No. R0694S) to remove any vector that did not receive an insert. The recombination products were then electroporated into electrocompetent cells (NEB 10-beta, New England BioLabs, cat. No. C3020K) and plated onto Carbenicillin plates. Sanger sequencing of 32 colonies, using n40.dn.F and EGFP.up.R primers (Supplementary Table S5), was then used to confirm the proper assembly of the library. The library was purified using a number of colonies needed to achieve the desired number of Barcodes (n = 50) to be associated at each sequence. Barcode-associated fragments were amplified using P7-pLSmp-ass-gfp (100 μM) and P5-pLSmP-ass-i741 primers (Supplementary Table S5), purified using Plasmid plus midi kit (QIAGEN) and tested for its quality via sequencing on a MiSeq (see below section).

MiSeq

The association between enhancer sequences and the barcodes was ascertained using Illumina MiSeq v2.0 sequencing with pair-ended 150 bp read length. The reads overlap for 30 bp in the middle of the enhancer sequences. Three MiSeq libraries were sequenced to obtain enough number of barcodes covering each tested enhancer sequence with the total number of 168 million reads. The barcodes with 15 bp length were sequenced at the same time with the same read names as the enhancer sequences with the same batch of MiSeq.

Lentivirus packaging and MOI

The lentiMPRA library was packaged into lentivirus using the plasmid library, psPAX2 (RRID: Addgene_12260) and pMD2.G (RRID: Addgene_12259) and its titration was determined as previously described[30]. In brief, iPSCs were plated at 0.150 million cells/well in 24-well plates and incubated for 24 h. Serial volumes (0, 2, 4, 8, 16, 32 ul) of the lentivirus were added. The infected cells were cultured for 3 days and washed with PBS three times before genomic DNA extraction. Genomic DNA was extracted using the Wizard SV genomic DNA purification kit (Promega). Virus titer and copy number of viral particles per cell were measured by qPCR as previously described[30].

iPSCs reprogramming and maintenance

Two iPSC lines were used for the lentiMPRA experiment: ACE1815 (two technical replicates, named ACE_a and ACE_b in the figures) and 11251. Lines were generated from human skin fibroblasts obtained from a skin biopsy of normal individuals using a viral-free episomal reprogramming method[57,61] at the Yale Stem Cell Reprogramming Core and passaged for 18 or 20 passages before differentiation. All iPSC lines used in this study fulfilled standard reprogramming criteria, including (i) immunocytochemical expression of pluripotency markers (NANOG; SSEA4; TRA1-60); (ii) expression of known hESC/iPSC markers (SOX2, NANOG, LIN-28, GDF3, OCT4, DNMT3B) by semi-quantitative RT-PCR; (iii) downregulation of exogenous reprograming factors. The iPSC lines were cultured on Matrigel (Corning Matrigel Matrix Basement Membrane Growth Factor Reduced)-coated dishes with mTeESR1 media (StemCell Technologies) and propagated using Dispase (StemCell Technologies).

Lentiviral infection

The lentiMPRA library was transduced into iPSCs at the undifferentiated stage. To provide 75%-80% confluency the next day, iPSCs were seeded on Matrigel-coated 10 cm dishes in mTeSR1 media supplemented with 10 uM Y-27632. After 24 h, the cells were infected with the lentivirus library at an average of Multiplicity of Infection (MOI) of 4–5 and incubated for 3 days with daily media changes to remove non-integrated virus. The experiment involved three independent replicate cultures, including two lines from different individuals (ACE1815 and 11251), and one technical replicate (ACE1815_a and ACE1815_b). All cultures were infected at the same time and using the same lentiviral library.

Forebrain organoid differentiation

The iPSC lines were differentiated into forebrain organoids as described in Jourdon et al. 2023[57] based on modifications of previous protocols[62,63]. Briefly, undifferentiated iPSC colonies were treated with 5 µM of the Y27632 compound and dissociated to single cells with Accutase (Millipore, 1:2 dilution in PBS 1X). Four million dissociated cells were seeded in each well of a 6-well plate and cultured in mTeSR1 with 10 µM Y27632 compound on an orbital shaker at a speed of 95 rpm. Forebrain neural induction was triggered by dual SMAD inhibition in mTeSR1 media supplemented with 10 µM SB431542, 1 µM LDN193189 and 5 µM Y-27632 (day1). At day 2, embryoid bodies were cultured in KSR medium (DMEM supplemented with 15% Knockout Serum Replacement, 1% L-Glutamine, 1% NEAA, 1% Pen/Strep and 55 µM of 2β-Mercaptoethanol, 2-ME) with the addition of SB431542, LDN193189, XAV939 and Y-27632. The neural induction with dual SMAD inhibition was maintained until day 7 after which organoids were gradually adapted to NIM medium (DMEM/F12, 1% N2 supplement, 2% B27 without vitamin A, 1% NEAA, 1% Pen/Strep, 0.15% Glucose and 1% Glutamax) through a dilution series of KSR and NIM. Neural progenitors proliferation was induced at day 9 in NIM 75% and KSR 25% supplemented with FGF2 (10 ng/ml) and EGF (10 ng/ml) and organoids were maintained in proliferative medium until day 16 in 100% NIM. Terminal differentiation was initiated at day 17 (TD0) in terminal differentiation medium (Neurobasal medium supplemented with 1% N2, 2% B27 without vitamin A, 15 mM HEPES, 1% Glutamax, 1% NEAA and 55 µM 2-ME) with the addition of the neurotrophic factors BDNF (10 ng/ml) and GDNF (10 ng/ml) until TD30. In the differentiation phase, half of the medium was changed twice a week. Organoids were transferred from wells of a 6-well plate to a 10 cm dish between TD5 and TD10 and the speed of the orbital shaker was decreased to 80 rpm.

Cell harvesting, library preparation and DNA/RNA extraction and sequencing for barcodes count

The infected cells were harvested at three different time points: as iPSCs, after 3 days from the infection before starting organoid differentiation (iPSC stage), and as forebrain organoids at earlier (TD0) and more mature (TD30) terminal differentiation points. Genomic DNA and total RNA were simultaneously extracted using the AllPrep DNA/RNA Mini Kit (Qiagen, cat. No. 80204) following the manufacturer's protocol. RNA samples were treated with Turbo DNase (Life Technologies, cat. No. AM1907) to remove any residual DNA contamination. Sequencing libraries were prepared as previously described[30]. Briefly, at least 60 µg total RNA per sample was used for reverse transcription with SuperScript II (Life Technologies, cat. No. 18064–071) using the primer P7-pLSmP-ass16UMI-gfp (Supplementary Table S5) to add a 16-bp UMI and a P7 flowcell sequence downstream of the barcode. PCR steps were performed on the DNA and RNA samples in order to amplify barcodes, adding P5 flowcell sequence and sample index upstream, and P7 flowcell sequence and UMI downstream to the barcode. Finally, the sequencing libraries were pooled and subjected to paired-end sequencing with UMI, and sample index read.

Immunostaining

Organoids were randomly selected and fixed in 4% PFA in PBS for 2–4 h. The organoids were then cryopreserved in 25% sucrose overnight, embedded in O.C.T. (Sakura), and frozen on dry ice before being stored at − 80 °C. Serial cryosections were obtained with a thickness of 12–16 µm. Immunostaining was performed by incubating the sections in blocking solution (PBS, 10% Donkey Serum, 1% Triton-100) for 1 h, followed by incubation with primary antibodies (overnight, 4 °C) and secondary antibodies (1–2 h, from Jackson ImmunoResearch or ThermoFisher Scientific). The slides were then mounted with coverslips using VECTASHIELD (Vector Labs) and imaged on a Zeiss microscope equipped with an apotome module and ZEN 3.3 (ZEN pro) software. Three cell lines were used for immunocytochemical analyses, and a minimum of four organoids per line were analyzed. Images were acquired randomly to cover the entire extent of the organoid. Antibody list: FOXG1 (rabbit, 1:200, Takara), PAX6 (mouse, 1:200, BD Bioscience), EOMES (rabbit, 1:1000, Abcam), FOXP2 (goat, 1:200, Santa Cruz), GAD1 (mouse, 1:200, Chemicon), HuC/D (mouse, 1:200, Invitrogen), SOX1 (goat, 1:100, R&D Systems), CTIP2 (rat, 1:500, Abcam).

For quantification, 4 organoids per sample were analyzed. Images were acquired from multiple sections to cover the entire extent of the organoid. Quantification of the average number of PAX6-, FOXG1- and CTIP2-positive cells was performed using Fiji software with the BioVoxxel plug-in under the Fiji analysis software platform. The relative number of PAX6 + , FOXG1 + or CTIP2 + cells was calculated as a percentage of total DAPI + cells across the 4 organoids in each iPSC line/time point.

LentiMPRA computational pipeline

Pre-processing using MPRAflow and MPRAnalyze

The association between enhancer sequences and barcodes was identified using the MPRAflow association package[30] with the following command:

```
nextflow run association.nf –fastq-insert "R1.fastq.gz" –fastq-bc "barcode.fastq.gz" –design "design.fa" –name "MPRAflow" –fastq-insertPE "R2.fastq.gz" -w {workdir} –labels "labels.txt" –cigar 270 M.
```

The count for number of reads for each RNA and DNA barcodes in all 9 samples for all enhancers were calculated using MPRAflow count package[30] with the following command:

```
nextflow run count.nf –dir "DNA_RNA" –e "experiments.csv" –design "design.fa" –association "MPRAflow_filtered_coords_to_barcodes.pickle" –labels "labels.txt" –umi-length 15 –name "countUMI" –outdir {workdir} –mpranalyze.
```

MPRAnalyze[64] was then performed to normalize the count for RNA barcodes and DNA barcodes in all 9 samples to generate the enhancer activity (RNA/DNA ratio) using the standard pipeline.

Identification of negative control distribution using negative references

A mixed Gaussian distribution was applied to calculate the distribution of negative controls using the 135 negative control activity in 9 samples. The curve fit function in Scipy package[65] was applied to estimate the mean and standard deviation of the negative control distribution. The distribution accurately described the negative control data with Kolmogorov–Smirnov test p-value = 0.33 and Anderson–Darling test p-value = 0.12.

Candidate regulatory sequences activity quantification and validation

The p-value for the activity of an enhancer to be higher than the negative control distribution was calculated using the enhancer activity (RNA/DNA) and the negative control distribution. The significantly active enhancers were then identified using the p-value of 0.05 after Bonferroni correction using the total number of enhancers tested.

Estimating sensitivity of MPRA experiment to define an enhancer as an active one

Using the negative control Gaussian distribution, we defined a cutoff for the RNA/DNA ratio that corresponds to a significant p-value. We then calculated the area under the positive control Gaussian and above the cutoff. The sensitivity was then estimated as a ratio of the area and the total area under the positive control Gaussian distribution.

Clustering of enhancer activity for candidate regulatory sequences

Enhancer activity (RNA/DNA) of candidate regulatory sequences and positive reference sequences was calculated by MPRAnalyze. Subsequently, Seaborn 0.11.0[66] clustermap was employed for clustering the enhancer activity values. Ward's minimum variance method was utilized for clustering. The resulting heatmap displayed the clustered enhancer activity values. The color used for enhancer activity larger than 4 was the same as that used for the value of 4.

Transcription factor binding site identification and expression of the transcription factors

FIMO[40] was employed to identify TFBSs from all 7289 sequences, including candidate regulatory sequences, positive reference sequences, and negative reference sequences. The transcription factor binding site annotation was downloaded from JASPAR2022[67] with 1252 Homo Sapiens annotations. The command for running FIMO[40] is as follows:

```
fimo –o {out_dir} JASPAR2022_tfbs input.fa.
```

TFBSs identified by FIMO with FDR < =0.05 were taken for further analyses. We quantified the number of TFBSs in each sequence and compared the distribution between validated and non-validated candidate regulatory sequences using Kolmogorov–Smirnov test. We further calculated the expression (RPKM) of the TFs which bind to these TFBSs in 97 TD0/TD30 organoid samples. The expression of TFs in validated and non-validated candidate regulatory sequences was compared using the Kolmogorov–Smirnov test.

Genes regulated by candidate regulatory sequences and expression

Genes regulated by candidate enhancers were identified by Amiri et al.[31] taking confident_set1, confident_set2, and proximity genes. The expression (RPKM) of the genes was calculated from 97 TD0/TD30 organoid samples.

Intersection with external datasets

Intersection with external datasets were performed using BedTools[68] and the resulting data was plotted using the 'UpSetR' package in R[69].

Single cell ATAC-seq analysis

For each 10X scATACseq sample, fastq files were first processed by cellranger-atac v2.0.0 with default parameters and 10X prebuilt reference arc-GRCh38-2020-A-2.0.0. The resulted cell-by-peak count matrix was first processed and filtered by Signac following online vignette (https://stuartlab.org/signac/articles/pbmc_vignette.html). Cell types were annotated using our annotated scRNAseq dataset and the label transfer method implemented in Seurat following online vignette (https://satijalab.org/seurat/articles/atacseq_integration_vignette.html). scATACseq peaks were then called using all reads in a sample or subsets of reads from each annotated cell type by running Signac function CallPeaks with default parameters.

Data availability

The source MPRA data described in this manuscript are available via the PsychENCODE Knowledge Portal (https://psychencode.synapse.org/). The PsychENCODE Knowledge Portal is a platform for accessing data, analyses, and tools generated through grants funded by the National Institute of Mental Health (NIMH)

PsychENCODE Consortium. Data is available for general research use according to the following requirements for data access and data attribution: (https://psychencode.synapse.org/DataAccess). For access to content described in this manuscript see: https://doi.org/https://doi.org/10.7303/syn51296303. The source bulk RNA-seq data are available at the NIMH Data Archive (NDA) under collection #C2424, at url: https://nda.nih.gov/edit_collection.html?id=2424.

Received: 8 September 2023; Accepted: 11 February 2024
Published online: 16 February 2024

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308. https://doi.org/10.1016/0092-8674(81)90413-x (1981).
2. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 16. https://doi.org/10.1186/s13072-015-0009-5 (2015).
3. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286. https://doi.org/10.1038/nrg3682 (2014).
4. Crawford, G. E. *et al.* DNase-chip: A high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3**, 503–509. https://doi.org/10.1038/nmeth888 (2006).
5. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218. https://doi.org/10.1038/nmeth.2688 (2013).
6. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49. https://doi.org/10.1038/nature09906 (2011).
7. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216. https://doi.org/10.1038/nmeth.1906 (2012).
8. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187. https://doi.org/10.1038/nature09033 (2010).
9. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384. https://doi.org/10.1371/journal.pbio.1000384 (2010).
10. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293. https://doi.org/10.1126/science.1181369 (2009).
11. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021 (2014).
12. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98. https://doi.org/10.1016/j.cell.2011.12.014 (2012).
13. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64. https://doi.org/10.1038/nature08497 (2009).
14. Krietenstein, N. *et al.* Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554-565 e557. https://doi.org/10.1016/j.molcel.2020.03.003 (2020).
15. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502. https://doi.org/10.1038/nature05295 (2006).
16. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**, 185–192. https://doi.org/10.1016/j.ygeno.2015.06.007 (2015).
17. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077. https://doi.org/10.1126/science.1232542 (2013).
18. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164. https://doi.org/10.1016/j.ygeno.2015.06.005 (2015).
19. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277. https://doi.org/10.1038/nbt.2137 (2012).
20. Mulvey, B., Lagunas, T. Jr. & Dougherty, J. D. Massively parallel reporter assays: Defining functional psychiatric genetic variants across biological contexts. *Biol. Psychiatry* **89**, 76–89. https://doi.org/10.1016/j.biopsych.2020.06.011 (2021).
21. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **172**, 1132–1134. https://doi.org/10.1016/j.cell.2018.02.021 (2018).
22. Kreimer, A. *et al.* Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* **13**, 1504. https://doi.org/10.1038/s41467-022-28659-0 (2022).
23. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* **25**, 713-727 e710. https://doi.org/10.1016/j.stem.2019.09.010 (2019).
24. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U. S. A.* https://doi.org/10.1073/pnas.2007049118 (2021).
25. Whalen, S. *et al.* Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron* **111**, 857-873 e858. https://doi.org/10.1016/j.neuron.2022.12.026 (2023).
26. Girskis, K. M. *et al.* Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**, 3239-3251 e3237. https://doi.org/10.1016/j.neuron.2021.08.005 (2021).
27. Rummel, C. K. *et al.* Massively parallel functional dissection of schizophrenia-associated noncoding genetic variants. *Cell* **186**, 5165-5182 e5133. https://doi.org/10.1016/j.cell.2023.09.015 (2023).
28. Guo, M. G. *et al.* Integrative analyses highlight functional regulatory variants associated with neuropsychiatric diseases. *Nat. Genet.* **55**, 1876–1891. https://doi.org/10.1038/s41588-023-01533-5 (2023).
29. Deng, C. *et al.* Massively parallel characterization of psychiatric disorder-associated and cell-type-specific regulatory elements in the developing human cortex. *bioRxiv* (2023). https://doi.org/10.1101/2023.02.15.528663
30. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412. https://doi.org/10.1038/s41596-020-0333-5 (2020).
31. Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* https://doi.org/10.1126/science.aat6720 (2018).
32. Barakat, T. S. *et al.* Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell* **23**, 276-288 e278. https://doi.org/10.1016/j.stem.2018.06.014 (2018).
33. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88-92. https://doi.org/10.1093/nar/gkl822 (2007).
34. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858. https://doi.org/10.1038/nature07730 (2009).

35. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330. https://doi.org/10.1038/nature14248 (2015).
36. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251. https://doi.org/10.1038/s41586-020-2559-3 (2020).
37. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461. https://doi.org/10.1038/nature12787 (2014).
38. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82. https://doi.org/10.1038/nature11232 (2012).
39. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. https://doi.org/10.1038/nature11247 (2012).
40. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064 (2011).
41. Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* https://doi.org/10.1126/science.aay1645 (2020).
42. Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213. https://doi.org/10.1038/s41586-021-03209-8 (2021).
43. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* https://doi.org/10.1038/s41586-022-05279-8 (2022).
44. Gupta, N. *et al.* Stress granule-associated protein G3BP2 regulates breast tumor initiation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1033–1038. https://doi.org/10.1073/pnas.1525387114 (2017).
45. Singh, G. *et al.* A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Res.* **31**, 564–575. https://doi.org/10.1101/gr.272468.120 (2021).
46. Mikdache, A. *et al.* Rgs4 is a regulator of mTOR activity required for motoneuron axon outgrowth and neuronal development in zebrafish. *Sci. Rep.* **11**, 13338. https://doi.org/10.1038/s41598-021-92758-z (2021).
47. Gallagher, D. *et al.* Ankrd11 is a chromatin regulator involved in autism that is essential for neural development. *Dev. Cell* **32**, 31–42. https://doi.org/10.1016/j.devcel.2014.11.031 (2015).
48. Vullhorst, D. *et al.* Selective expression of ErbB4 in interneurons, but not pyramidal cells, of the rodent hippocampus. *J. Neurosci.* **29**, 12255–12264. https://doi.org/10.1523/JNEUROSCI.2454-09.2009 (2009).
49. Nakajima, H. & Koizumi, K. Family with sequence similarity 107: A family of stress responsive small proteins with diverse functions in cancer and the nervous system (Review). *Biomed. Rep.* **2**, 321–325. https://doi.org/10.3892/br.2014.243 (2014).
50. Montilla, A., Mata, G. P., Matute, C. & Domercq, M. Contribution of P2X4 receptors to CNS function and pathophysiology. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21155562 (2020).
51. Platzer, K. *et al.* Haploinsufficiency of CUX1 causes nonsyndromic global developmental delay with possible catch-up development. *Ann. Neurol.* **84**, 200–207. https://doi.org/10.1002/ana.25278 (2018).
52. Cubelos, B. *et al.* Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* **66**, 523–535. https://doi.org/10.1016/j.neuron.2010.04.038 (2010).
53. Nieto, M. *et al.* Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers II-IV of the cerebral cortex. *J. Comp. Neurol.* **479**, 168–180 (2004).
54. Edri, R. *et al.* Analysing human neural stem cell ontogeny by consecutive isolation of Notch active neural progenitors. *Nat. Commun.* **6**, 6500. https://doi.org/10.1038/ncomms7500 (2015).
55. Dennis, D. J. *et al.* Neurog2 and Ascl1 together regulate a postmitotic derepression circuit to govern laminar fate specification in the murine neocortex. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4934–E4943. https://doi.org/10.1073/pnas.1701495114 (2017).
56. Herbst, F. *et al.* Extensive methylation of promoter sequences silences lentiviral transgene expression during stem cell differentiation in vivo. *Mol. Ther.* **20**, 1014–1021. https://doi.org/10.1038/mt.2012.46 (2012).
57. Jourdon, A. *et al.* Modeling idiopathic autism in forebrain organoids reveals an imbalance of excitatory cortical neuron subtypes during early neurogenesis. *Nat. Neurosci.* **26**, 1505–1515. https://doi.org/10.1038/s41593-023-01399-0 (2023).
58. Lalanne, J.-B. *et al. Multiplex profiling of developmental enhancers with quantitative, single-cell expression reporters* (BioRxiv, 2023).
59. Zhao, S. *et al.* A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat. Genet.* **55**, 346–354. https://doi.org/10.1038/s41588-022-01278-7 (2023).
60. Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895–908. https://doi.org/10.1016/j.cell.2012.12.041 (2013).
61. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412. https://doi.org/10.1038/nmeth.1591 (2011).
62. Kadoshima, T. *et al.* Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell-derived neocortex. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20284–20289. https://doi.org/10.1073/pnas.1315710110 (2013).
63. Watanabe, K. *et al.* Directed differentiation of telencephalic precursors from embryonic stem cells. *Nat. Neurosci.* **8**, 288–296 (2005).
64. Ashuach, T. *et al.* MPRAnalyze: Statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183. https://doi.org/10.1186/s13059-019-1787-z (2019).
65. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).
66. Waskom, M. L. Seaborn: Statistical data visualization. *J. Open Source Softw.* https://doi.org/10.21105/joss.03021 (2021).
67. Castro-Mondragon, J. A. *et al.* JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173. https://doi.org/10.1093/nar/gkab1113 (2022).
68. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. https://doi.org/10.1093/bioinformatics/btq033 (2010).
69. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364 (2017).

Acknowledgements

We thank the members of the Vaccarino lab for extensive discussions, technical help and contributions to methods. We thank Guilin Wang and Christopher Castaldi, and the Yale Center for Genome Analysis for library preparation, deep sequencing, and Cell Ranger analysis. We thank Caihong Qiu and Jason Thomson at the Yale Stem Cell Center for the generation of the iPSC lines. Part of the illustrations were created using BioRender.Com. Data were generated as part of the PsychENCODE Consortium. See Supplemental Information or visit https://doi.org/10.7303/syn24240356 for a complete list of grants and PIs.

Author contributions

F.M.V. and A.A. conceived the study, designed, and supervised experiments; N.A. and F.I. contributed to experimental design and provided essential reagents; D.C. prepared the MPRA library, generated organoid preps and processed them for bulk RNA-seq and DNA-seq; Y.W., F.W., D.C. and S.N. performed bioinformatic analyses; D.C. and J.M. performed organoid immunocytochemical analyses; D.C. and Y.W. generated display items and wrote the manuscript; all authors provided edits and comments on the manuscript.

Funding

We acknowledge the following grant support: National Institute of Mental Health R01 MH109648 (FMV), R56 MH114911 (FMV), R56 MH114899 (AA), R56 MH114901 (GC), U01 MH116438 (NA), the Simons Foundation awards No. 632742 (FMV, AA).

Competing interests

NA is the cofounder and on the scientific advisory board of Regel Therapeutics and receives funding from Bio-Marin Pharmaceutical Incorporated. All other authors do not hold any competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-54302-7.

Correspondence and requests for materials should be addressed to A.A. or F.M.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.