TITLE: A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of genetic variants *in vivo*

AUTHORS:
Tomas Lagunas Jr.[1,2,3], Stephen P. Plassmeyer[1,2], Ryan Z. Friedman[1,3], Michael A. Rieger[1,2,3], Anthony D. Fischer[1,2], Alessandra F. Aguilar Lucero[4], Joon-Yong An[5,6], Stephan J. Sanders[4], Barak A. Cohen[1], Joseph D. Dougherty[1,2]

[1]Department of Genetics, Washington University School of Medicine, 660 S. Euclid Ave, Saint Louis MO, 63108, USA.
[2]Department of Psychiatry, Washington University School of Medicine.
[3]Division of Biology and Biomedical Sciences, Washington University School of Medicine.
[4]Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neuroscience, University of California, San Francisco, CA 94518
[5]Department of Integrated Biomedical and Life Science, Korea University, Seoul 02841, Republic of Korea
[6]School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul 02841, Republic of Korea

Contact Information:
Dr. Joseph Dougherty
Department of Genetics
Campus Box 8232
4566 Scott Ave.
St. Louis, MO. 63110-1093
P: 314-286-0752
F: 314-362-7855
E:jdougherty@wustl.edu

Acknowledgements:
We'd like to thank Bernie Mulvey, Dana King, and the Djuranovic lab for discussions and advice. We'd also like to thank Kristian Quigless, Christian Doss, as well as Mingje Li and the Hope Center Viral Vectors Core for technical support, as well as the CGS spike-in cooperative (especially Jess Hoisington-Lopez and MariaLynn Crosby), and GTAC@MGI for sequencing support. This work was funded by the Simons Foundation (571009) and the NIH (5R01MH116999) and T32 (MH014677, GM007067).

ABSTRACT:

Human genetic studies have identified a large number of disease-associated *de novo* variants in presumptive regulatory regions of the genome that pose a challenge for interpretation of their effects: the impact of regulatory variants is highly dependent on the cellular context, and thus for psychiatric diseases these would ideally be studied in neurons in a living brain. Furthermore, for both common and rare variants, it is expected that only a subset fraction will affect gene expression. Massively Parallel Reporter Assays (MPRAs) are molecular genetic tools that enable functional screening of hundreds of predefined sequences in a single experiment. These assays have been used for functional screening of several different types of regulatory sequences *in vitro*. However, they have not yet been adapted to query specific cell types *in vivo* in a complex tissue like the mouse brain. Here, using a test-case 3′UTR MPRA library with variants from ASD patients, we sought to develop a method to achieve reproducible measurements of variant effects *in vivo* in a cell type-specific manner. We implemented a Cre-dependent design to control expression of our library and first validated our system *in vitro*. Next, we measured the effect of >500 3′UTR variants in excitatory neurons in the mouse brain. Finally, we report >40 variants with significant effects on transcript abundance in the context of the brain. This new technique should enable robust, functional annotation of genetic variants in the cellular contexts most relevant to psychiatric disease.

INTRODUCTION:

In the current era of common and rare variant genome-wide approaches, thousands of candidate genetic variants with potential association to psychiatric and neurological diseases have been uncovered, the vast majority in noncoding, presumably regulatory, DNA regions. For common variants, large collaborative studies have identified dozens of genomic regions that are significantly associated with disease (Consortium et al., 2020; Grove et al., 2019; Matoba et al., 2020), but each region contains hundreds to thousands of noncoding variants, of which only a subset are thought to have a functional consequence and potentially be causal. For rare variants, whole-genome sequencing has identified thousands of noncoding variants per individual, and efforts at associating these with disease would benefit from knowing which can indeed alter gene expression. However, in either case, defining specific functional variants has proven to be a major challenge given the large number that need to be screened. Furthermore, cell-type context plays an important role in gene-regulation studies (Mulvey et al., 2020). For example, neurons express a variety of neuron-specific transcription factors (e.g. (Hevner et al., 2006)) and RNA-binding proteins (Pilaz and Silver, 2015) as they mature, and thus genomic variants that alter binding sites for these would only show effects in mature neurons. Therefore, there is a need for a high-throughput method that can be easily adapted to functionally screen these genetic variants in a parallel fashion, specifically in the cellular contexts relevant to diseases of the central nervous system (CNS). For most psychiatric diseases, this ideal cellular context would be specific classes of neurons, *in vivo.*

One example of a set of variants that would be of great interest to test in *vivo* would be *de novo* mutations discovered in individuals with Autism Spectrum Disorder (ASD). In the past decade, numerous *de novo* mutations have been directly implicated in ASD (Satterstrom et al., 2020; Werling et al., 2018). Initial analyses focused on mutations in coding regions, which are more readily interpreted for functional effects than noncoding variants (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012a, 2015) . However, there is estimated to be substantial additional burden from noncoding mutations (An et al., 2018; Turner et al., 2017). This can include both transcriptional regulators, like promoters and enhancers, as well as 5′/3′ untranslated regions (UTRs). UTRs contain several classes of regulatory elements that control mRNA stability, subcellular localization, and rate of translation for their cognate transcript (Mayr, 2017). The interpretation of noncoding variants presents several challenges. Firstly, they lack a simple triplet code, and their effects are likely to be highly cell-type dependent. Furthermore, an overwhelming number of these mutations are being discovered. (Werling et al., 2018) reported a total of 71,132 *de novo* noncoding variants from whole genome sequencing of 519 ASD families; of these, 737 fall in UTRs. Finally, based on experience with *de novo* variants in protein-coding regions, we expect any ASD risk mediated by *de novo* variants in UTRs to be from a small subset of variants with large effect sizes.

Massively Parallel Reporter Assays (MPRAs) are genetic tools that could address these challenges since they can be used to functionally assay several thousand predefined sequences at once (Mulvey et al., 2020). These assays have enabled functional annotation of

thousands of noncoding genomic elements, as well as the impact of variants in UTRs in particular, prioritizing potentially causal changes (Choi et al., 2020; Kircher et al., 2019; Litterman et al., 2019; Siegel et al., 2020). In addition, emerging MPRA studies have begun to dissect the role of 3′UTR variation (Griesemer et al., 2021) in function and regulatory activity *in vitro*. Unsurprisingly, there is only a modest overlap of functional elements across six diverse human cell lines, underscoring the density of elements with cell type-specific regulatory potential within UTRs. Furthermore, there are limits to the extent to which an *in vitro* system, even primary cells or iPSC derived neural systems, can recapitulate the normal gene expression and thus regulatory landscape seen during neuronal development *in vivo.* Thus, in the context of ASD and neuropsychiatric disease, elements would ideally be assayed in the brain and in relevant cell types in order to more accurately model the effect of these variants.

However, few studies have been successful in implementing MPRAs to measure the effect of genomic elements *in vivo* and, to date, none have done so in a cell-type specific manner. Several examples exist of MPRAs being applied *in vivo* in non-CNS tissues, such as in zebrafish embryos (Rabani et al., 2017; Smith et al., 2013), the mouse liver (Patwardhan et al., 2012), and mouse retina (Hughes et al., 2018; White et al., 2013). Yet, only three studies have implemented MPRAs in the brain. The first, (Shen et al., 2016) showed how packaging a complex library in AAV (the preferred method for transgene delivery to CNS), did not affect element representation. However, recovery of elements from the mouse brain after transduction proved to be very challenging — only a fraction of the total library of the 45,000 barcoded elements was recoverable and expression was highly variable between biological replicates. Other emerging studies (e.g., (Lambert et al., 2021)) have shown reproducible recovery of smaller AAV packaged libraries from the mouse cortex when screening for potential enhancers but did not assess the impact of genetic variants. Finally, another emerging study (Shen et al., 2019) successfully used multiplex reporter barcoding to detect the effects of a single variant in mouse cortical explants. This demonstrated that, at least for one variant, variant effects could be detected by MPRAs with sufficient coverage, suggesting larger scale *in vivo* studies of variants might be possible. However, a caveat to all studies thus far is that the libraries are assessed in a mix of cell types that are transduced. This could partially mask the effect of variants regulated by effectors that are differentially expressed across brain cell types. Thus, there is an urgent need for a system to assay the effects of noncoding variants at scale in a cell type-specific manner.

Here, we describe the development of a high-throughput cell-type specific MPRA approach for the mouse brain, with the sensitivity to measure the effects of individual variants. As a test-case we used a 3′UTR MPRA library to functionally assay several hundred *de novo* variants found in the genomes of ASD cases and sibling controls. We developed a Cre recombinase-dependent library design. We first piloted this in a mouse neuroblastoma cell line, assessing total RNA and RNA paired with a ribosome affinity purification to enable assessment of both transcriptional and translational effects. We then optimized the delivery of these same elements to cortical neurons *in vivo*. We were indeed able to assess the functional consequences of hundreds of variants in parallel. We found effects of variant alleles are highly cell type-specific, and report several mutations that substantially alter host transcript abundance in neurons in the developing brain,

and thus could possibly be functional in ASD. In all, the approach here should enable large-scale assessment of the functional impact of variants from psychiatric genetics in specific cell types in the brain.

MATERIALS AND METHODS:

Animal models

Veterinary care and housing is provided by the veterinarians and veterinary technicians of Washington University School of Medicine under Dougherty lab's approved IACUC protocol. All protocols involving animals were completed with: Tg(RBP4-cre) KL100Gsat/Mmcd (RRID:MMRRC_037128-UCD; Beltramo et al., 2013), Slc32a1tm2(cre)Lowl/J (catalog #16962, The Jackson Laboratory; RRID:IMSR_JAX:016962; Vong et al., 2011), and Vglut1-IRES2-Cre-D strain (Jackson Stock No: 023527). All mice were genotyped following a standard protocol of taking clipped toes into lysis buffer (0.5M Tris-HCl pH 8.8, 0.25M EDTA, 0.5% Tween-20, 4uL/mL of 600 U/mL Proteinase K enzyme) for 1 hour to overnight. This is followed by heat denaturation at 99 C for 10 minutes. 1 uL of the resulting lysate was used as a template for PCR with with 500 nM froward and reverse primers, using 1x Quickload Taq Mastermix (NEB) with the following cycling conditions: 94 1 min, (94 30 sec, 60 30 sec, 68 30 sec) x 30 cycles, 68 5 min, 10 hold.

MPRA plasmid library preparation

For non-Cre-dependent reporter expression, we used a previously described pmrPTRE-AAV backbone which contained the following elements: CMV promoter, T7 promoter, mtdTomato CDS, hGH terminator, and flanking ITRs. The T7 promoter and mtdTomato CDS were amplified from pmrPTRE-AAV using PTRE_floxed_F/R and Phusion High-Fidelity PCR Master Mix (NEB). NotI and SalI sites added by the primers were used to subclone this amplicon into pRM1506_TMM432. The final pmrPTRE-floxed-AAV backbone consists of a floxed cassette containing the T7 promoter and tdTomato CDS in reverse orientation with respect to a CAG promoter, followed by a bGH terminator, all flanked by ITRs.

The oligo sequences designed for this library are provided in **[Supp Table 4]**. The UTR contexts for each oligo were taken from GRh37/hg19 by centering a maximum 120 bp window around the variant position. Variant allele sequences were substituted at the reference position to generate the alternative allele UTR context. For indels, the UTR context was limited to the minimum context that would fit either allele, and padding sequences were added outside of cloning cut sites. Additional elements with known or suspected post-transcriptional regulatory roles were included as well: the alpha component of the WHP posttranscriptional regulatory element (WPRE) and synthetic elements consisting of four tandem sequences for either the Smaug response element (SRE), Pumilio response element (PRE), or Quaking response element (QRE).

A constant 20 bp linker sequence separates the UTR context from a nine bp barcode sequence. Each UTR context was repeated in the design with six unique barcodes. Barcodes were selected to be Hamming distance of two apart and to exclude cut sites and homopolymers longer than three bases. Priming sites and cut sites were added to both ends to generate 210 bp oligos which were synthesized by Agilent Technologies.

The synthesized oligos were amplified with 4 cycles of PCR using Phusion polymerase and primers Bactin_FWD/REV. Amplicons were PAGE purified and digested with NheI and KpnI (NEB). Library inserts were cloned into pmrPTRE-floxed-AAV with T4 ligase (Enzymatics) and transformed into chemically competent DH5α (NEB). Outgrowths were plated on LB agar plates with 100 µg/mL carbenicillin, and approximately 71,000 colonies were counted, allowing us to capture the entire design at 95% confidence, assuming a 50% synthesis error rate. Plates were scraped, and the collected pellets were cultured for an additional 12 hours in LB with carbenicillin before preparing glycerol stocks and maxi preps (Qiagen).

Cell culture

Mouse neuroblastoma N2a cells were maintained at 5% CO2 37°C, and 95% relative humidity in DMEM (Gibco) supplemented with 10% fetal bovine serum (FBS, Atlanta Biologicals). Human neuroblastoma SH-SY5Y cells were maintained similarly, except with DMEM/F12 (Gibco) substituted as the basal medium. Cells were also incubated with 1% penicillin-streptomycin (Gibco). For transient transfections, antibiotics were excluded from the transfection medium and re-introduced upon media change 12 hours post-transfection. Cells were passaged with 0.25% Trypsin-EDTA (Gibco) every 2-3 days or once they reached 80-90% confluency.

Cell culture TRAP

For each cell culture TRAP experiment, six replicate T75 flasks (TPP or Sarstedt) were seeded in advance with mouse N2a neuroblastoma cells to be 80-90% confident by the time of transfection. For the library cloned into the pmrPTRE-AAV backbone, 20 µg of total plasmid containing an equimolar ratio of reporter library and an Ef1a-EGFP-RPl10a construct was transfected. For the Cre-inducible library, 23 µg of total plasmid was transfected, consisting of equimolar ratios of the library, an DIO-EF1a-EGFP-RPl10a construct, and an Ef1a-Cre construct. Transient transfections were performed with Lipofectamine 2000 (Invitrogen), and DNA:lipid complexes were prepared by co-incubation in Opti-MEM I (Gibco) for 30 minutes prior to transfection. Transfection medium was replaced 12 hours following transfection, and cells were harvested for TRAP after an additional 24 to 36 hours.

TRAP was performed as described (Heiman et al., 2014) with minimal modification. Briefly, cells were incubated in 100µg/mL cycloheximide (Sigma) for 15 minutes at 37°C prior to harvest. Cells were rinsed twice with 5 mL of DMEM 100 µg/mL cycloheximide before being lifted into 5 mL of DMEM 100 µg/mL cycloheximide. Cells were pelleted by spinning at 500xg for 5 minutes at 4°C. The DMEM was replaced with 2 mL of ice-cold cell lysis buffer (10 mM pH 7.4 HEPES, 1% NP-40, 150 mM KCl, 10 mM MgCl2, 0.5 mM dithiothreitol, 100 µg/ml CHX, protease inhibitors, and RNase inhibitors) and cells were lysed on ice. Lysates were clarified by centrifugation at 2000xg for 10 minutes at 4°C. DHPC (Avanti) was added to a final concentration of 30mM, and lysates were incubated on ice for 5 minutes. A tenth of the volume was taken as the Input, and the remaining volume was incubated with protein L-conjugated magnetic beads (Invitrogen) coupled with a mixture of two monoclonal anti-GFP antibodies (Doyle et al., 2008). The beads were incubated for 4h at 4°C prior to four washes with a high-salt buffer (10 mM pH 7.4 HEPES, 1% NP-40, 350 mM KCl, 10 mM MgCl2, 0.5 mM dithiothreitol, 100 µg/ml CHX, protease inhibitors, and RNase inhibitors) before resuspension in cell-lysis buffer.

Input and TRAP RNA was extracted using Trizol LS (Life Technologies). Extracted RNA samples were DNase treated (Ambion) and cleaned by column-based purification (Zymo

Research). Concentrations and RNA quality were determined using RNA ScreenTapes and a 4200 Tapestation System (Agilent Technologies). All RINe measurements exceeded 9.

Parallel Plasmid DNA for each replicate was recovered from each cell pellet following lysis using the Qiagen DNeasy Blood & Tissue Kit, and prepared for sequencing in parallel to RNA, as below. We found that having multiple replicate DNA libraries was critical for reducing variance in element activity measurements, at the transcript abundance level in particular. As such we recommend preparation of replicate DNA libraries, either from the plasmid input or from recovered plasmid from each experimental replicate of transfected cells.

***in vivo* MPRA**

Two Vglut1-IRES2-Cre-D litters were subjected to intracranial injections for delivery of the library packaged in AAV9. P0-P2 pups were incubated on ice to anesthetize by inducing hypothermia for ~10 minutes. An aliquot of the MPRA library packaged in AAV9 (~10^9 vg/uL) was drawn up in 33G Hamilton syringe with a 1 mm needle. Pups were brought up to the needle and 1 uL of virus was injected at three positions per brain hemisphere hemisphere (6 total injections per pup). Pups were taken directly to the warming pad until pups fully recovered (~20 minutes). After recovery, pups were placed back into the cage with the mother and monitored every 24 hours for one week. At P21 brains were harvested for extraction of RNA.

We aimed to determine the source of this jackpotting and reasoned either the barcodes were all present in the starting template of total RNA and our library preparation was not efficient at a particular step, or the barcodes were simply too low abundance in the starting RNA pool. To this end, we conducted a series of technical replicates splitting a sample at each step of the library preparation protocol: cDNA synthesis, cDNA amplification, adapter ligation, and indexing PCR **[Supp Fig 4A]**. Taking a single RNA sample and doing two separate cDNA synthesis reactions for independent sequencing libraries resulted in jackpotting (PCC < 0.4) **[Supp Fig 4B]**. Taking cDNA from a single sample and amplifying it in two independent reactions for library preparation also led to jackpotted samples (PCC < 0.4) **[Supp Fig 4C]**. However, if the amplified cDNA from a single sample was taken into two independent reactions for adapter ligation, then the final sequencing libraries were highly correlated (PCC > 0.9) **[Supp Fig 4D]**. This was the case for reactions split at the final indexing PCR as well (PCC > 0.9) **[Supp Fig 4E]**. This result revealed to us that the source of jackpotting is at the cDNA synthesis or amplification steps. To investigate this further, we employed a variety of techniques that included reaction splitting/repooling to boost scale, unique molecular identifiers (UMIs), and emulsion PCR (ePCR). Reaction splitting/repooling and ePCR did not alleviate any of the jackpotting issues at any of the stages tested (data not shown). We then incorporated UMIs at the cDNA synthesis step in order to precisely quantify and eliminate PCR duplicates. After sequencing the resulting libraries and computationally collapsing UMIs, we found that only a small fraction of elements was recoverable. Together, these results led us to conclude that this jackpotting was, in fact, a representation of the barcodes present in the RNA: for a given amount (100 ng) of RNA from the brain, relatively few barcode molecules were present. Consistent with this, increasing input RNA up to 1 ug reduced jackpotting effects, but still resulted in relatively low sample correlations (PCC < 0.4).

MPRA sequencing library preparation

Libraries were prepared by taking total RNA or TRAP RNA and performing cDNA synthesis using Superscript III Reverse Transcriptase standard protocol with pmrPTRE_floxed_AAV_antisense (GCATAAAAAACAGACTACATAATACTG) for library specific priming. Resulting cDNA or plasmid DNA, were then used for PCR to amplify libraries using

Phusion polymerase (Thermo) using library specific primers pmrPTRE_AAV_sense (GCATGGACGAGCTGTACAAG) and pmrPTRE_floxed_AAV_antisense. Reactions were purified using AMPure XP beads between each step. The purified PCR products were then digested with NheI and KpnI restriction enzymes for 1 hour at 37 deg C. The purified digested products were ligated to 4 equimolar staggered adapters (this is to provide sequence diversity for sequencing). Ligated products were purified and then used for a second PCR using Illumina primers for library indexing. The purified libraries were then QC'ed and subjected to quality control and then 2x150 next generation sequencing on an Illumina NovaSeq.

BC counting and normalization

Sequencing reads were trimmed using cutadapt v1.16 (Martin, 2011) and aligned to the library reference sequences using bowtie2 v2.3.5 (Langmead and Salzberg, 2012) using "very sensitive" settings. Barcodes were counted from aligned reads with mapping quality of 10 or greater using a custom Python script. Counts within each sample were normalized to each sequencing library size using edgeR (Robinson et al., 2010) as counts per million (CPM) prior to downstream analysis.

The abundance of each element in RNA samples were normalized to their abundance either the transfected plasmid or transduced viral libraries by averaging CPM across barcodes within each sample. These averages were divided by the average for each element across barcodes in the respective DNA library. The $\log_2$ of these ratios was taken as the expression for each element. Similarly, for ribosomal occupancy and translation efficiency per element was calculated by normalizing TRAP RNA to DNA counts and TRAP RNA to Input RNA counts, respectively.

Element filtering and differential expression analysis

A paired Student's t-test was performed to test for a difference between the mean element-wise expression of the alternative and reference alleles within each biological replicate, using the t.test function in R. Before testing, thresholds for element inclusion were determined by a grid search of count, barcode number, and replicate number thresholds that maximized the number of variants significant at a Benjamini-Hochberg FDR < 0.05. Briefly, at increasing count thresholds, variants within each replicate were retained if both alleles had more than a given threshold for barcodes above said count threshold, and variants with both alleles passing count and subsequently barcode thresholds in a minimum number of replicates were selected for analysis. For the *in vitro* MPRA, variants must be present with both alleles having three barcodes with at least 10 counts from both RNA and DNA in all six replicates. For the *in vivo* MPRA, variants must be present with both alleles having three barcodes with at least 200 counts from both RNA and DNA in four replicates.

T-test p-values for the variants passing the specified thresholds were corrected for multiple comparisons by using the p.adjust function in R to apply the Benjamini-Hochberg procedure for false discovery rate. The full set of results for these tests are provided in Supplemental Table 1 and 2.

Modelling allele by sex interations

Sex-differential allelic effects were examined in the *in vivo* MPRA using a linear mixed model for the interaction of these two terms. The same set of variants that passed count, barcode, and replicate thresholds for the differential expression analysis were included in this analysis. For each variant, replicate-wise barcode-averaged transcript abundances were fit to an allele by sex

interaction with random intercepts for biological replicate using the lmer package in R for linear mixed models, using the formula Expression ~ Allele * Sex + (1 | Replicate). The significance of each model coefficient was determined by likelihood ratio test (LRT) against a reduced model for each term. LRT p-values for each term were corrected for multiple comparisons using the p.adjust function to apply the Benjamini-Hochberg FDR procedure. Variants found to have a significant main effect of sex or a significant allele by sex interaction were confirmed using an appropriate non-parametric test, either a Wilcoxon signed-rank test for main effect comparisons or a Kruskal-Wallis test for allele by sex interactions.

Fluorescent immunohistochemistry and analysis

Brains were harvested from postnatal day 21 mice, and one hemisphere was chosen for subsequent RNA extraction/TRAP. The remaining hemisphere was fixed for 48 h in 4% paraformaldehyde followed by 24 h in 15% sucrose in 1× PBS and then 24 h in 30% sucrose in 1× PBS. The hemisphere was then frozen in OCT compound (optimum cutting temperature compound; catalog #23-730-571, Thermo Fisher Scientific). A Leica CM1950 cryostat was used to create 40 µm sagittal sections of brain tissue. Sections were immediately placed in a 12-well plate containing 1X PBS and 0.1% w/v sodium azide.
For immunostaining, sections were incubated in a blocking solution (1× PBS, 5% donkey serum, 0.25% Triton-X 100) for 1 h in a 12 well plate at room temperature, then with rabbit anti-RFP primary antibody (1:500; Rockland catalog #600-401-379) in blocking solution overnight in a sealed 12 well plate at 4°C. Following three five-minute washes in PBS, sections were incubated in donkey anti-rabbit Alexa Fluor 568 secondary antibody (1:1000, Invitrogen catalog #A10042) and DAPI (in blocking solution for 1 h. Sections were washed as before, and during the second wash, 1 µg/mL DAPI was added. Sections were slide mounted with Prolong Gold and visualized for anti-RFP and DAPI staining on a Zeiss Axio Imager Z2 four-color inverted confocal microscope. TdTomato-positive cells were quantified by hand using FIJI (Schindelin et al., 2012).

Machine learning

Gapped k-mer SVM models were fit using gkmSVM (Ghandi et al., 2014) with the parameters -l 4 -k 4 -m 1 (4-mers) and -l 5 -k 5 -m 1 (5-mers). Stratified five-fold cross-validation and computing ROC and PR curves was performed using scikit-learn version 0.19.1 (Pedregosa et al., 2011).

RESULTS:

Cre-dependent MPRA reproducibly measures allelic effects in a mouse neuroblastoma cell line

As a proof of principle for an allelic effect MPRA, we examined *de novo* variants identified within annotated 3′UTRs from the whole-genome sequencing of 519 families from the Simons Simplex Collection (Werling et al., 2018), totaling 342 mutations from probands and 299 from unaffected siblings within the same cohort. For each variant we synthesized an allelic pair of 3′UTR ‘elements’ spanning 120 bp of sequence centered on the variant. To be able to compare biological to non-biological sequences, for 322 variants, we randomly shuffled the sequence to generate a set of GC-matched controls. We tagged all 1,624 elements with six unique barcodes to provide internal replicates and control for potential barcode effects. To enable eventual cell-

type specific studies, we cloned the final library of 9,744 synthesized oligos into the 3′UTR of a membrane-localized tdTomato reporter embedded in a Double-floxed inverse Orientation (DiO) cassette (Schnütgen et al., 2003), such that the reporter library would only express following Cre-mediated recombination **[Fig 1A-B]**.

To first evaluate whether our assay could detect variant effects on reporter transcript abundance and translation, we co-transfected the library into mouse neuroblastoma N2a cells with two additional constructs–one expressing Cre recombinase, and another expressing eGFP-tagged large ribosomal subunit protein L10a (eGFP-RPL10a). The eGFP-RPL10a construct allows us to employ the Translating Ribosome Affinity Purification (TRAP) technique to measure the effects of UTR elements on ribosome occupancy (Heiman et al., 2008). We harvested RNA from six replicate transfections from both the whole-cell lysate (Input) and the polysome-bound TRAP fraction (Heiman et al., 2014) **[Fig 1C]**. Barcode sequencing libraries were prepared from both Input RNA and TRAP RNA**,** to identify elements that alter ribosome occupancy(TRAP) on top of effects on transcript abundance(Input). We also conducted DNAseq on the plasmid DNA re-extracted from the transfected cells to enable normalization of each RNA barcode to its starting abundance in the cells.

We examined the coverage and reproducibility of the assay, and the range of the biological activity across elements. We sequenced to an average depth of 5,388 counts per barcode. In the DNA, 8,053 barcodes had non-zero counts, suggesting a <20% element dropout at the cloning stage. Cloning efficiency correlated with element GC-content, as elements with less than 40% GC content cloned less efficiently **[Supp Fig 1A]**. A corresponding 85% of elements were represented with at least three barcodes and carried forward for analysis **[Supp Fig 1B]**. In the RNA data, correlations of barcode abundance between replicate libraries from both Input and TRAP generally exceeded 0.99 (PCC) **[Fig 2A]**, indicating high reproducibility. Correlations of either RNA measure with barcode abundance in recovered plasmid libraries averaged 0.96 (PCC), indicating that variation in reporter abundance was largely driven by DNA copy number, as the range of differences in cloning efficiency exceeds the magnitude of expected biological effects of elements. Thus, we normalized input RNA counts to plasmid DNA counts for subsequent analyses. This revealed variation in steady state RNA abundance across elements, with 99% of elements spanning -1.33 to 0.95 $\log_2$-normalized expression (RNA/DNA) **[Fig 2B],** indicating that the sampled UTR elements exhibit a modest 5-fold range in transcript abundance as measured by our assay.

Normalizing TRAP RNA abundance by DNA copy number revealed a similar dynamic in the ribosomal occupancy of reporter transcripts. However, these differences are driven primarily by the underlying difference in transcript abundance. Normalizing TRAP RNA abundance to the input RNA abundance a proxy for 'Translation Efficiency'; TE, defined here as $\log_2$ TRAP/Input counts), showed a narrow dynamic range from -0.40 to 0.45, indicating 3′UTR effects on ribosome occupancy are more subtle than on reporter transcript abundance. Interestingly, pairwise comparison of genome-derived reference elements to GC-matched shuffled control sequences showed that random sequences had both lower transcript abundance (Wilcoxon signed-rank $p = 4.28 \times 10^{-6}$) and TE ($p = 3.05 \times 10^{-5}$) than their corresponding reference sequences **[Fig 2B]**. This suggests that genomic sequences generally promote higher steady-state transcript abundance than random sequences. However, the elements containing *de novo* variants (alternative alleles; Alt) did not show a systematic difference from their paired reference allele (Ref) elements. This is not unexpected, as most are small or single base mutations, and only a small subset of human mutations, even from probands, might be presumed to be strongly functional *a priori*.

Biological effects should be driven by specific sequence elements in the UTRs, and thus activity should be somewhat predictable from primary sequence. To establish a biological signature of active elements, we trained k-mer support vector machines (SVMs) (Ghandi et al., 2014) to classify the 200 highest-expressing elements from the 200 lowest-expressing elements, pooling Ref and shuffled sequences (Shuf). In this framework, each sequence is represented by the frequency of all possible k-mers as input to the SVM. We trained 4- and 5-mer SVMs with 5-fold cross-validation. To ensure the SVM was not overfit, we also fit SVMs on the same sequences with random labels. The SVMs achieved an area under the receiver operating characteristic (AUROC) of 0.709-0.712 and an area under the precision recall curve (AUPRC) of 0.688-0.708 **[Fig 2C-D]**, while models fit on random labels could not classify the data (AUROC 0.512-0.518) **[Supp Fig. 2]**, indicating there are sequence-specific elements underlying UTR activity. To understand which sequences mediated these effects, we next scored all possible 4-mers against the 4-mer SVM. 4-mers predicted to be highly active tended to be GC rich, while 4-mers predicted to be inactive tended to be AT rich. We also used DREME (Bailey, 2011) to identify *de novo* motifs enriched in the high expressing sequences relative to the low expressing sequences and obtained similar results. Taken together, these results indicate a substantial fraction of the activity of UTRs is driven by sequence features captured by small motifs, and identifies the motifs with activity in N2a cells.

While more highly expressed elements tended to be GC rich, genomic elements were clearly different from random GC matched controls. Comparing each Ref element to its matched Shuf control revealed that 165 were significantly different (Benjamini-Hochberg FDR <0.05) with a median 1.2-fold change in expression **[Fig 2E]**. Thus, genomic sequences produce a specific level of activity upon which allelic effects are expected to act. Of the 257 tested comparisons, 63 showed a significant difference at a set threshold (25%) for change in expression. Of the significant changes, 48 were downregulating. Assuming equal probability of up- and down-regulation, this is more than expected by chance (hypergeometric p=0.00387, OR = 2.01), again reflecting the relative greater propensity for genomic derived UTR tiles to enhance steady-state reporter expression.

Finally, we examined allelic effects on steady-state transcript abundance. Of the designed variants, 519 met thresholds for barcode representation for both the Ref and Alt sequence and were included in the analysis. Of these, nearly half 251 (48.3%) showed significant (FDR < 0.05) Log_2 Fold Change of RNA/DNA (LFC), though mostly with very small effects **[Supp Table 1]**. This indicates that our assay has high sensitivity and suggests it is plausible nearly any mutation will have some effect, though for most the effect sizes are so small any biological impact would be very subtle. However, 31 (5.9%) of significant variants did show an absolute LFC corresponding to >25%, suggesting that a subset of UTR variants may have enough impact on transcript abundance **[Fig 2E]** to have measurable biological consequences. Contrasting effects on transcript abundance with ribosome occupancy again revealed that variant effects on TE tended to be much smaller, and no significant allelic effects survived multiple testing correction. Overall, our cell line assay confirmed reproducibility and robustness of our DiO 3′UTR MPRA design, motivating applying the approach to specific cell types *in vivo*.

Cre-dependent MPRA reproducibly measures functional effects of several hundred variants in excitatory neurons in the mouse brain

To assess the effect of these elements *in vivo,* the entire element library was packaged in adeno-associated virus serotype 9 (AAV9) for delivery into the mouse brain. We have previously shown (Cammack et al., 2020) that with AAV delivery we get widespread viral transduction in

the neocortex and mainly target neurons and astrocytes with serotype 9. We found that packaging of the library did not drastically change the range of distribution or barcode recovery rates and correlated well (PCC > 0.8) with the plasmid counts **[Supp Fig 3]**. Thus, packaging had no adverse effects on the composition of the library and moved forward with delivery *in vivo*.

Bioinformatic analysis of the expression patterns of genes associated with ASD have revealed a correlation structure of two loose modules - a module enriched for chromatin regulators with peak expression in immature excitatory neurons, and a module of synaptic-related proteins, with peak expression during critical periods of postnatal synaptogenesis and pruning (De Rubeis et al., 2014; Parikshak et al., 2013; Satterstrom et al., 2020; Willsey et al., 2013). Therefore, we first attempted to deliver the library to two neuronal sub-types, layer V pyramidal neurons and GABAergic interneurons, during this pruning period by using *Rbp4* and *Vgat* Cre drive lines, respectively. However, we discovered that only a small fraction of the delivered elements was recovered, and representation of barcodes was highly distorted and, in many cases, favoring a small, distinct subset in each biological replicate, resulting in low correlation between replicates (PCC < 0.2). We reasoned that since *Rbp4-positive* and *Vgat*-positive cells made up a small population of cells in the mouse brain, their low contribution to total cortical RNA may be a strong contributing factor to library jackpotting (See methods).

Therefore, we delivered the AAV library to a well-characterized excitatory neuron specific Cre line (Vglut1-IRES2-Cre-D (Harris, et al. 2014); Vlgut1^{Cre}) **[Fig 3A]**, which makes up a larger population of cells, covering all pyramidal cells of the cortex. We first confirmed the expression of the library by immunofluorescence **[Fig 3B]**. We saw widespread expression of the tdTomato reporter in cells with the morphology of pyramidal neurons with the perinatal injection yielding transductions across cortex (~3% of cells in cortex successfully targeted. See Methods for calculation). Importantly, Cre negative littermates showed no expression of the library, confirming cell-type specificity **[Fig 3C]**. Next, an additional 12 animals' cortices were collected for RNA. We sequenced, in all, 12 RNA replicates and 2 replicates of viral prep DNA to obtain RNA barcode and DNA barcode counts, respectively.

Next, we performed a similar quality control analysis as for N2a data above. Correlations of barcode abundance between biological replicates on average exceeded 0.80 (PCC) **[Fig 3D]**. Notably, this observed correlation is lower than our *in vitro* test, but increased variability is commensurate with lower rates of element delivery and recovery from a subset of cells in complex tissue. This increased variance motivated our doubling of the number of replicates to preserve statistical power. Similar to what was done for the N2a data, we removed elements which were absent in the DNA counts and filtered for a minimum sequencing depth and barcode number, resulting in 402 analyzed elements. Pairwise comparison of genome-derived Ref elements to GC-matched Shuf control sequences again showed that Shuf sequences had lower transcript abundance (Wilcoxon signed-rank $p = 4.28 \times 10^{-6}$) than their corresponding Ref sequences, as observed in N2as **[Fig 3G]**. Of the 190 testable Ref-Shuf comparisons, 78 showed a significant difference in expression. We also observed an overrepresentation of steady-state downregulation in shuffled elements (63) compared to their paired reference (hypergeometric p = 0.00073, OR = 2.034) **[Fig 3H]** FInally, we again used k-mer SVMs to determine if there were sequence features that predicted *in vivo* activity and achieved an AUROC of 0.676-0.680 and an AUPRC of 0.674-0.677 **[Fig 3E-F]**, comparable to the SVMs trained on *in vitro* activity.

We then assessed the impact of the 402 mutated alleles on transcript abundance for each element. Overall, we found that 41 (10.2%) showed significant (FDR < 0.05) changes **[Supp**

Table 2], and 35 (8.7%) were significant with a median absolute LFC corresponding to >25% change in expression **[Fig 3H]**. In both Ref vs. Shuf and Ref vs. Alt contrasts, compared to our N2a experiments, we found that there were larger effect sizes in the *in vivo* studies. We suspect that with additional replicates or higher transduced cell numbers we might also detect the numerous smaller fold changes as identified in N2a cells. Nonetheless, this demonstrated our ability to simultaneously assess the activity of hundreds of UTR elements and the functional impact of variants in a disease relevant cellular regulatory context *in vivo.*

In addition to the functional demonstration of allelic effects that, because of neuron-specific factors, might only be present *in vivo*, live animal assays could also enable detection of allelic effects dependent on other biological phenomena. For example, males are 4 times more likely than females to be diagnosed with ASD (Werling and Geschwind, 2013), suggesting some genetic effects may be sex-specific, a phenomena which cannot be fully modeled in cell lines.To examine the possiblity of sex-differential effects of particular *de novo* variants, we fit linear mixed-effect models for the interaction of allele and sex. Two variants, both from ASD cases, found in HOXC11 and ABHD2 showed sex-differntial effects after multiple tests correction **[Fig 4A]**. A post-hoc nonparametric test confirmed this interaction for the HOXC11 variant which appears to decrease expression in males but not females (Kruskal-Wallis p = 0.00373) **[Fig 4B]**. This highlights the power of this approach to study the impact of whole-organismal contexts like sex or environmental risk factors on gene regulation in a cell-type specific manner.

Finally, we were interested in comparing data sets from our two distinct contexts in order to dissect similarities and differences. In total, we examined all 402 elements passing QC in both Vglut and N2a. Of the 256 (N2a) and 41 (Vglut) significant variants, 19 were present in both data sets. However, focusing on those showing a >25% change in gene expression threshold, there was no overlap. This finding is unsurprising given the low correlation of expression values across the two systems **[Fig 5A].** Transcript abundance spanned a broader range in the *in vivo* assay (Brown-Forsythe $p < 2.2x10^{-16}$), highlighting the possibility that a more complex regulatory environment may contribute to a greater dynamic range **[Fig 5B]**. Furthermore, the cross-validated SVM scores of the N2a activity are uncorrelated to the observed activity in excitatory neurons **[Fig 5C]**, suggesting there are cell type-specific factors regulating UTR activity through interaction with specific sequences. This highlights the need to assess the function of noncoding variants in multiple contexts, and especially focusing on contexts where noncoding variants for a specific disease are most likely to act.

DiO 3'UTR MPRA reveals patient-derived mutations that alter transcript abundance

De novo protein-coding mutations, mostly heterozygous loss-of-function (LOF), are thought to account for 5-10% of ASD (Iossifov et al., 2014; Sanders et al., 2012b). Thus, *a priori* we would expect only a small fraction of patients would have *de novo* noncoding mutations driving their disease as well. Further, even for those that may carry a causal noncoding mutation, of the ~70 *de novo* noncoding variants per individual, only a very small subset are expected to mediate disease risk. Thus, our goal was to develop an approach to screen hundreds of variants in a complex context like the mouse brain and in a cell-type specific manner. As our proof-of-principle, we focused on UTR variants because some studies had indicated increased rates in UTR mutations in ASD cases (Turner et al., 2017), and because UTRs can be more readily linked to individual genes than noncoding elements such as enhancers. Here, we have successfully identified UTR variants as functional across two contexts — in an neuroblastoma cell-line and in excitatory neurons in the mouse brain. Any potentially causal variants are likely among the very small subset with larger effect sizes. If the mutations identified in the coding sequence are a guide, for many genes a 50% reduction (heterozygous LOF mutation) can cause disease. We saw few alleles with this effect size (~2), but did find several dozen case

mutations with a more moderate effect size of >25% across the two systems **[Fig. 6 A-B]**. We validated a subset of these with Sanger sequencing from the patient samples, and confirmed 29 were indeed bona fide *de novo* mutations, with a validation rate of 100% for any fragment that was amplified **[Supp Table 3]**. Since at this cutoff, rates of mutations were not higher in cases than in controls (N2a: 17/31 from probands, hypergeometric p = 0.4, odds ratio = 1.04, Neurons: 21/35 from probands, hypergeometric p = 0.4, odds ratio = 1.13), we would expect only a very small subset of these, at most, would be causal. Nonetheless this screen has identified several high impact alleles of interest for further investigation.

DISCUSSION:

Here we describe the development of a cell-type specific *in vivo* MPRA. We demonstrate that the method is sensitive enough to identify allelic effects for hundreds of variants in parallel, and we provide a proof-of-concept analysis of several hundred 3′UTR mutations from ASD probands and their unaffected siblings. This approach should be directly applicable to the thousands more UTR variants already discovered in psychiatric disease genetic studies, and readily adaptable to assaying noncoding variants found in other relevant positions such as promoters and enhancers. Additionally, it should also be usable for dissecting the sequence dependence of previously identified regulatory elements with activity in neurons, using MPRA libraries designed for saturation mutagenesis of potential binding motifs (Kircher et al., 2019) and other perturbations (Rieger et al., 2020). Thus, the approach should have both translational and basic science applications.

This work and its challenges allowed us to deeply characterize the range of conditions, from environment to sequence context, that influence these regulatory assays. Our first attempts in delivering this library to rarer cell types using Rbp4 and Vgat Cre-lines were limited by low element recovery rates, making reproducible measurement of many variants in parallel intractable. Careful analysis of all stages of RNAseq library prep revealed jackpotting originated at cDNA synthesis, suggesting reporter mRNA was diluted beyond the point of efficient recovery. This is consistent with the relative sparseness of GABAergic cortical neurons compared to cortical excitatory neurons, indicating they will contribute less to the total RNA of the cortex. Use of neither emulsion PCR, reaction splitting, nor UMI incorporation in second strand synthesis could resolve this fundamental limitation. However, when we delivered to a more abundant cell type, increasing the barcode concentration in the final total RNA, the jackpotting was largely resolved. We do note that variability *in vivo* with AAV was still higher than when delivering to N2as in culture (PCC of >0.8 vs >0.9) with transfection, where delivering to >70% of cells at high copy number is straightforward. However, we were able to overcome this increased variability by increasing sample number. For *in vivo* assays, what other approaches might work to allow access to these rarer cell types and overcome the low barcode abundance in the starting RNA? Three general approaches come to mind: targeting AAV delivery to hit a larger portion of the Cre-positive cells (for example, adult injections into regions where GABAergic neurons are a larger fraction, such as the striatum), reducing the complexity of the library (using a smaller number of total barcodes, making each barcode more likely to be well represented), or enriching for the barcoded RNAs prior to cDNA synthesis, either by a targeted capture of reporter RNAs, or potentially purifying the Cre-positive cells by FACS or TRAP. Any of these might further expand the current method to rarer Cre populations. Nonetheless, the current iteration of the technology already enables access to assessing variants in the regulatory context of mature neurons, an essential cell type for many CNS diseases.

In all, we also discovered dozens of variants that altered transcript abundance in these cells. Since 3′UTRs are frequently bound by regulators such as miRNAs and RBPs, we suspected that many of the variants assayed here would have their impact post-transcriptionally (e.g., on RNA decay or translation efficiency). Thus, future studies are needed to determine if the variants uncovered here are acting via altering transcriptional or post-transcriptional regulation. As some aspects of RNA surveillance depend on RBPs that are loaded onto transcripts during splicing (Lykke-Andersen and Jensen, 2015), it may also be of interest to determine how the impact of 3′UTR variants might be further unmasked in reporter libraries that contain introns. Likewise, upstream open reading frames (uORFs) in 5′UTRs can also regulate ribosome loading and translational efficiency, and thus it will also be of interest to test patient mutations in 5′UTRs, as well as the interaction between reporter uORFs and 3′UTR mutations.

We were also interested in the extent to which our *in vivo* MPRA might inform the genetic architecture of ASD, as some studies have revealed a trend for enrichment of mutations in UTRs in ASD cases when compared to their siblings (OR 1.1, nominal p-value p<.04)(Turner et al., 2017), providing statistical evidence that a subset are causal. We therefore tested whether subsetting to those variants that had functional effects (>25% change and FDR<.05) in either N2as or neurons would show enrichment in cases over controls. At these cut-offs, we saw no enrichment, but it is possible that focusing on the variants with even larger effect sizes, and/or potentially those occurring in the constrained genes that are often impacted in forms of ASD mediated by *de novo* coding mutations, might reveal an enrichment of case mutations over controls. Such enrichment might allow estimates of what fraction of ASD could be caused by functional UTR mutations overall, though we need to screen a much larger number of variants to perform a robust burden test. Nonetheless, several of the functional variants identified here are found in known (NRXN1 (Gauthier et al., 2011)) and plausible (RBFOX2 (Partridge and Carter, 2017), HMGB1 (Dipasquale et al., 2017)) ASD genes.

One surprising discovery was that variants resulted in a roughly equal number of up- and down-regulating events relative to the reference allele. Most sequencing studies of *de novo* mutations in ASD have focused on heterozygous stop-gain mutations that are clearly loss-of-function and predicted to reduce protein expression by half. We would expect that mutations that have similar magnitude decreases in RNA via disrupting transcript abundance in the same genes might lead to ASD. However, the presence of mutations that increase RNA abundance highlight the fact that a different class of genes might be involved in causing ASD by increased level rather than decreased level. Such genes might overlap with ASD genes found in ASD-associated duplications such as 7q11.23 (Sanders et al., 2015), or protein gain-of-function mutations in channel genes such as those that cause some epileptic syndromes (Miceli et al., 2015). Scaling the approach to the thousands of variants uncovered in the most recent ASD whole genome sequencing should provide a statistical signal if such categories of UTR mutations contribute to disease.

Finally, outside of UTRs and ASD, we envision the use of cell type-specific MPRAs *in vivo* for identification of functional variants across several different diseases and in other noncoding regions of the genome. Furthermore, these methods present unique opportunities to perform these regulatory assays in the most relevant and specific biological context for a given disease. Altogether, we anticipate these methods will aid in the study of noncoding disease risk and inspire new adaptations of MPRAs.

FIGURES:

Fig 1. ASD 3′UTR library design and delivery.
A) MPRA library constructs were designed with a CAG promoter (prom) driving the TdTomato reporter, followed by the 3′UTR oligo reference or alternative sequence (with or without variant, respectively) that is uniquely barcoded. B) All elements were uniquely barcoded 6 times. Cloning of this library was completed in the Double-floxed inverse Orientation (DiO) design for Cre-dependent expression. C) MPRA library, Cre recombinase and TRAP allele were delivered via plasmid transfection into N2a cells. Following incubation, total RNA and TRAP RNA were isolated to prepare sequencing libraries to then count BCs to calculate expression per element.

Fig 2. Screen in mouse neuroblastoma cell line identifies variants that alter steady state transcript abundance.
A) Scatter plots showing correlation between replicates of RNA vs RNA (left), RNA vs plasmid DNA (center), and RNA vs TRAP (right) RNA CPM counts (in order from left to right). B) Pairwise comparison of expression value distribution among Ref, Alt, and Shuf sequences in Transcript Abundance, Ribosomal Occupancy, and Translation Efficiency data sets. Significance is denoted by asterisk. C) ROC and D) PRC curves for k-mer SVMs to classify high and low expressing elements. Shaded area represents 1 standard deviation based on five-fold cross-validation E) Volcano plot for Ref vs Shuf elements (purple) in library showing significance (y-axis) vs log2 FC (x-axis) overlaid with volcano plot for Alt vs Ref elements (green). Horizontal dashed line corresponds to FDR 0.05 and vertical dashed lines correspond to log2 FC equivalent to 25% change in expression.

Fig 3. Screen in excitatory neurons in the mouse brain identifies variants that alter steady state transcript abundance.
A) MPRA library was packaged into AAV9 and delivered into perinatal mouse cortices via intracranial injection and later harvested at P21 for RNA extraction. B) Immunofluorescence of P21 brain showing localization of tdTomato expression (from MPRA library) in excitatory neurons with nuclei labeled with DAPI (blue). C) Immunofluorescence of Cre-negative littermate showing DAPI staining and no signal from the Cre-dependent MPRA library. D) Scatter plot showing correlation between RNA counts of biological replicates. E) ROC and F)PRC curves for k-mer SVMs to classify high and low expressing elements G) pairwise comparison of Alt, Ref, and Shuf sequence expression. H) Volcano plot for Ref vs Shuf elements (purple) in library showing significance (y-axis) vs log2 FC (x-axis) overlaid with volcano plot for Alt vs Ref elements (green). Horizontal dashed line corresponds to FDR 0.05 and vertical dashed lines correspond to log2 FC equivalent to 25% change in expression.

Figure 4. In vivo assay captures potential sex dependence of variants
A) Volcano plot showing significance vs log2 FC of elements in library when considering sex effects and using a more complete linear model that includes allele only, sex only, or allele:sex effects (in order from left to right). Horizontal dashed line corresponds to FDR 0.05. B) Allele expression differences, by sex, for ABDH2 and HOXC11. Significance denoted by asterisk.

Figure 5. Low correlation and differences in range of expression between in vitro and in vivo suggest importance of regulatory context
A) Scatter plot of in vitro vs in vivo expression values B) Comparison of in vitro and in vivo expression distributions, Brown-Forsythe leve-type test for difference in variance $p < 2.2e-16$ C) in vitro SVM predictions vs in vivo expression

Figure 6. Reproducible measurements in mouse neuroblastoma cell line and excitatory neurons in the mouse brain
Relative transcript abundance log2 (RNA/DNA) for reference and alternative alleles from A) N2a and B) Vglut experiments.

Supplemental Figure 1. Library quality control
A) Average element abundance vs element GC content B) Barcode recovery by sample

Supplemental Figure 2. SVM data from random labels
A) ROC and B)PRC for k-mer SVMs to classify high and low expressing shuffled elements

Supplemental Figure 3. Viral packaging correlates with plasmid DNA barcode counts
A) Scatter plot showing correlation between plasmid DNA and viral DNA barcode counts

Supplementary Figure 4: Reaction splitting to determine source of jackpotting
A) Various steps of MPRA library preparation pipeline. B) Library correlation of technical replicates when splitting at the cDNA synthesis stage. C) Library correlation of technical replicates when splitting at the cDNA amplification stage. D) Library correlation of technical replicates when splitting at the adapter ligation stage. E) Library correlation of technical replicates when splitting at the indexing PCR stage.

Supplemental Table 1: All significant N2a elements with their corresponding LFC and significance value (p-val, FDR, Bonferroni).

Supplemental Table 2: All significant Vglut elements with their corresponding LFC and significance value (p-val, FDR, Bonferroni).

Supplemental Table 3: Variant validation from proband (p1) and *de novo* validation by absence in parents (fa/mo).

Supplemental Table 4: Oligo pool sequences included in the MPRA library and their corresponding barcodes.

REFERENCES:

1. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*.
2. Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.
3. Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. Proc. Natl. Acad. Sci. *117*, 10003–10014.
4. Choi, J., Zhang, T., Vu, A., Ablain, J., Makowski, M.M., Colli, L.M., Xu, M., Hennessey, R.C., Yin, J., Rothschild, H., et al. (2020). Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. Nat. Commun. *11*, 2718.
5. Consortium, T.S.W.G. of the P.G., Ripke, S., Walters, J.T., and O'Donovan, M.C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. MedRxiv 2020.09.12.20192922.
6. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.
7. Dipasquale, V., Cutrupi, M.C., Colavita, L., Manti, S., Cuppari, C., and Salpietro, C. (2017). Neuroinflammation in Autism Spectrum Disorders: the Role of High Mobility Group Box 1 Protein. Int. J. Mol. Cell. Med. IJMCM *6*, 148–155.
8. Gauthier, J., Siddiqui, T.J., Huashan, P., Yokomaku, D., Hamdan, F.F., Champagne, N., Lapointe, M., Spiegelman, D., Noreau, A., Lafrenière, R.G., et al. (2011). Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. Hum. Genet. *130*, 563–573.
9. Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLOS Comput. Biol. *10*, e1003711.
10. Griesemer, D., Xue, J.R., Reilly, S.K., Ulirsch, J.C., Kukreja, K., Davis, J., Kanai, M., Yang, D.K., Montgomery, S.B., Novina, C.D., et al. (2021). Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. BioRxiv 2021.01.13.424697.
11. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444.
12. Heiman, M., Schaefer, A., Gong, S., Peterson, J.D., Day, M., Ramsey, K.E., Suárez-Fariñas, M., Schwarz, C., Stephan, D.A., Surmeier, D.J., et al. (2008). A Translational

Profiling Approach for the Molecular Characterization of CNS Cell Types. Cell *135*, 738–748.
13. Heiman, M., Kulicke, R., Fenster, R.J., Greengard, P., and Heintz, N. (2014). Cell type–specific mRNA purification by translating ribosome affinity purification (TRAP). Nat. Protoc. *9*, 1282–1291.
14. Hevner, R.F., Hodge, R.D., Daza, R.A.M., and Englund, C. (2006). Transcription factors in glutamatergic neurogenesis: Conserved programs in neocortex, cerebellum, and adult hippocampus. Neurosci. Res. *55*, 223–233.
15. Hughes, A.E.O., Myers, C.A., and Corbo, J.C. (2018). A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. Genome Res. *28*, 1520–1531.
16. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y., Narzisi, G., Leotta, A., et al. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron *74*, 285–299.
17. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
18. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R.J.A., Costello, J.F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. *10*, 3583.
19. Lambert, J.T., Su-Feher, L., Cichewicz, K., Warren, T.L., Zdilar, I., Wang, Y., Lim, K.J., Haigh, J., Morse, S.J., Canales, C.P., et al. (2021). Parallel functional testing identifies enhancers active in early postnatal mouse brain. BioRxiv 2021.01.15.426772.
20. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
21. Litterman, A.J., Kageyama, R., Tonqueze, O.L., Zhao, W., Gagnon, J.D., Goodarzi, H., Erle, D.J., and Ansel, K.M. (2019). A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. Genome Res.
22. Lykke-Andersen, S., and Jensen, T.H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat. Rev. Mol. Cell Biol. *16*, 665–677.
23. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.
24. Matoba, N., Liang, D., Sun, H., Aygün, N., McAfee, J.C., Davis, J.E., Raffield, L.M., Qian, H., Piven, J., Li, Y., et al. (2020). Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. Transl. Psychiatry *10*, 1–14.
25. Mayr, C. (2017). Regulation by 3′-Untranslated Regions. Annu. Rev. Genet. *51*, 171–194.
26. Miceli, F., Soldovieri, M.V., Ambrosino, P., Maria, M.D., Migliore, M., Migliore, R., and Taglialatela, M. (2015). Early-Onset Epileptic Encephalopathy Caused by Gain-of-Function Mutations in the Voltage Sensor of Kv7.2 and Kv7.3 Potassium Channel Subunits. J. Neurosci. *35*, 3782–3793.
27. Mulvey, B., Lagunas, T., and Dougherty, J.D. (2020). Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants Across Biological Contexts. Biol. Psychiatry.
28. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. Nature *485*, 242–245.
29. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly

interconnected protein network of de novo mutations. Nature *485*, 246–250.

30. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell *155*, 1008–1021.
31. Partridge, L.M.M., and Carter, D.A. (2017). Novel Rbfox2 isoforms associated with alternative exon usage in rat cortex and suprachiasmatic nucleus. Sci. Rep. *7*, 9929.
32. Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.
33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
34. Pilaz, L.-J., and Silver, D.L. (2015). Post-transcriptional regulation in corticogenesis: how RNA-binding proteins help build the brain. Wiley Interdiscip. Rev. RNA *6*, 501–515.
35. Rabani, M., Pieper, L., Chew, G.-L., and Schier, A.F. (2017). A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. Mol. Cell *68*, 1083-1094.e5.
36. Rieger, M.A., King, D.M., Crosby, H., Liu, Y., Cohen, B.A., and Dougherty, J.D. (2020). CLIP and Massively Parallel Functional Analysis of CELF6 Reveal a Role in Destabilizing Synaptic Gene mRNAs through Interaction with 3′ UTR Elements. Cell Rep. *33*, 108531.
37. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
38. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012a). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.
39. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012b). *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.
40. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233.
41. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e23.
42. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.
43. Schnütgen, F., Doerflinger, N., Calléja, C., Wendling, O., Chambon, P., and Ghyselinck, N.B. (2003). A directional strategy for monitoring Cre-mediated recombination at the cellular level in the mouse. Nat. Biotechnol. *21*, 562–565.
44. Shen, S.Q., Myers, C.A., Hughes, A.E.O., Byrne, L.C., Flannery, J.G., and Corbo, J.C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res. *26*, 238–255.
45. Shen, S.Q., Kim-Han, J.S., Cheng, L., Xu, D., Gokcumen, O., Hughes, A.E.O., Myers, C.A., and Corbo, J.C. (2019). A candidate causal variant underlying both higher

intelligence and increased risk of bipolar disorder. BioRxiv 580258.

46. Siegel, D.A., Tonqueze, O.L., Biton, A., Zaitlen, N., and Erle, D.J. (2020). Massively Parallel Analysis of Human 3′ UTRs Reveals that AU-Rich Element Length and Registration Predict mRNA Destabilization. BioRxiv 2020.02.12.945063.
47. Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.
48. Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., et al. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. Cell *171*, 710-722.e12.
49. Werling, D.M., and Geschwind, D.H. (2013). Sex differences in autism spectrum disorders. Curr. Opin. Neurol. *26*, 146–153.
50. Werling, D.M., Brand, H., An, J.-Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet. *50*, 727–736.
51. White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. *110*, 11952–11957.
52. Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., et al. (2013). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. Cell *155*