




## A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of non-coding elements in vivo

Tomas Lagunas Jr.<sup>1,2,3</sup>, Stephen P. Plassmeyer<sup>1,2</sup>, Anthony D. Fischer<sup>1,2</sup>, Ryan Z. Friedman <sup>1,3</sup>, Michael A. Rieger<sup>1,2,3</sup>, Din Selmanovic<sup>1,2,3</sup>, Simona Sarafinovska <sup>1,2</sup>, Yvette K. Sol<sup>1,2</sup>, Michael J. Kasper<sup>1,2</sup>, Stuart B. Fass<sup>1,2</sup>, Alessandra F. Aguilar Lucero<sup>4</sup>, Joon-Yong An <sup>5,6</sup>, Stephan J. Sanders <sup>4</sup>, Barak A. Cohen <sup>1</sup> & Joseph D. Dougherty <sup>1,2</sup> 

The function of regulatory elements is highly dependent on the cellular context, and thus for understanding the function of elements associated with psychiatric diseases these would ideally be studied in neurons in a living brain. 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 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 genomic elements containing variants from autism patients, we developed a method to achieve reproducible measurements of element effects in vivo in a cell type-specific manner, using excitatory cortical neurons and striatal medium spiny neurons as test cases. This targeted technique should enable robust, functional annotation of genetic elements in the cellular contexts most relevant to psychiatric disease.

<sup>1</sup>Department of Genetics, Washington University School of Medicine, 660 S. Euclid Ave, Saint Louis, MO 63108, USA. <sup>2</sup>Department of Psychiatry, Washington University School of Medicine, 660 S. Euclid Ave, Saint Louis, MO 63108, USA. <sup>3</sup>Division of Biology and Biomedical Sciences, Washington University School of Medicine, 660 S. Euclid Ave, Saint Louis, MO 63108, USA. <sup>4</sup>Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neuroscience, University of California San Francisco, San Francisco, CA 94518, USA. <sup>5</sup>Department of Integrated Biomedical and Life Science, Korea University, Seoul 02841, Republic of Korea. <sup>6</sup>School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul 02841, Republic of Korea. ✉email: [jdougherty@wustl.edu](mailto:jdougherty@wustl.edu)

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 elements. For common variants, large collaborative studies have identified dozens of genomic regions that are significantly associated with disease<sup>1–3</sup>, but each region contains hundreds to thousands of elements containing 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 are found in elements that control gene expression in the brain, and thus might alter neuronal function. However, in either case, defining the effect of DNA elements 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<sup>4</sup>. For example, as they mature, neurons express a variety of neuron-specific transcription factors (TFs) (e.g., in ref. <sup>5</sup>) and RNA-binding proteins (RBPs)<sup>6</sup>, and thus elements containing their 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 elements 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*.

In the past decade, numerous *de novo* mutations have been directly implicated in autism<sup>7,8</sup>. Initial analyses focused on mutations in coding regions, which are more readily interpreted for functional effects than noncoding variants<sup>9–13</sup>. However, there is estimated to be substantial additional burden from noncoding mutations<sup>14,15</sup>. 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<sup>16</sup>. However, these regions pose challenges to study for functional effects since they don't follow a triplet code and are not easily interpretable.

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<sup>4</sup>. 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<sup>17–20</sup>. In addition, recent MPRA studies have begun to dissect the role of 3'UTR variation<sup>21</sup> 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 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.

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 elements, using a Cre recombinase-dependent library design. As a test-case, we used a 3'UTR MPRA library to functionally assay several hundred elements containing *de novo* variants found in the genomes of autism cases and sibling controls. We first piloted this in a mouse neuroblastoma cell line, assessing total RNA and RNA paired

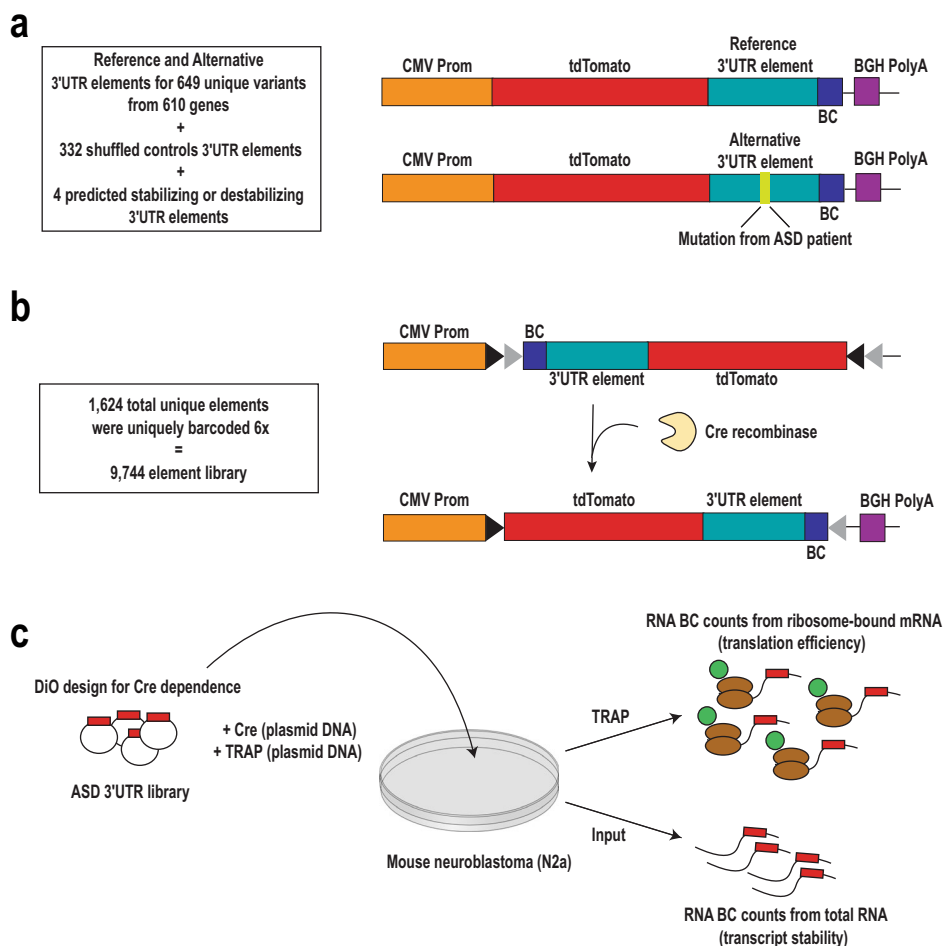
with a ribosome affinity purification to enable assessment of both transcriptional and translational effects. We then optimized the delivery of these same elements to two types of neurons *in vivo*. We were indeed able to assess the functional effects of hundreds of elements in parallel, and found effects of elements are highly cell type-specific. We also examined the ability to test for variant effects, and power calculations indicate this should be possible, but will require more extensive barcoding than used here. In all, the approach here should enable future large-scale assessment of the functional impact of variants from psychiatric genetics in specific cell types in the brain.

## Results

**Cre-dependent MPRA reproducibly measures element effects in a mouse neuroblastoma cell line.** As a proof of principle, we examined *de novo* variants identified within annotated 3'UTRs from the whole-genome sequencing of 519 families with autism, primarily from the Simons Simplex Collection<sup>8</sup>, targeting 342 mutations from probands and 307 from unaffected siblings within the same cohort (649 unique variants [Supplementary Data 1] to make 1298 ref/alt pairs). For each variant we synthesized an allelic pair of 3'UTR nucleotide stretches spanning 120 bp of sequence centered on the variant, which we term elements. To be able to compare biological to non-biological sequence elements, for 322 variants, we randomly shuffled the sequence to generate a set of GC-matched controls. Additionally, we included 4 predicted stabilizing/destabilizing controls. We tagged all 1624 elements with six unique barcodes to provide internal replicates and be able to measure potential for barcode effects. To enable eventual cell-type specific studies, we cloned the final library of 9744 synthesized oligos into the 3'UTR of a membrane-localized tdTomato reporter embedded in a Double-floxed inverse Orientation (DiO) cassette<sup>22</sup>, such that the reporter library would only express following Cre-mediated recombination [Fig. 1a, b].

To first evaluate whether our assay could detect UTR element 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<sup>23</sup>. We harvested RNA from six replicate transfections from both the whole-cell lysate (Input) and the polysome-bound TRAP fraction<sup>24</sup> [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 5388 counts per barcode. In the DNA, 8053 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 [Supplementary Fig. 1a]. A corresponding 85% of elements were represented with at least three barcodes and carried forward for analysis [Supplementary Fig. 1b]. In the RNA data, correlations of barcode abundance between replicate libraries from both Input and TRAP generally exceeded 0.99 (Pearson's Correlation Coefficient, PCC) [Fig. 2a], indicating high reproducibility (read depth, correlations, etc, for



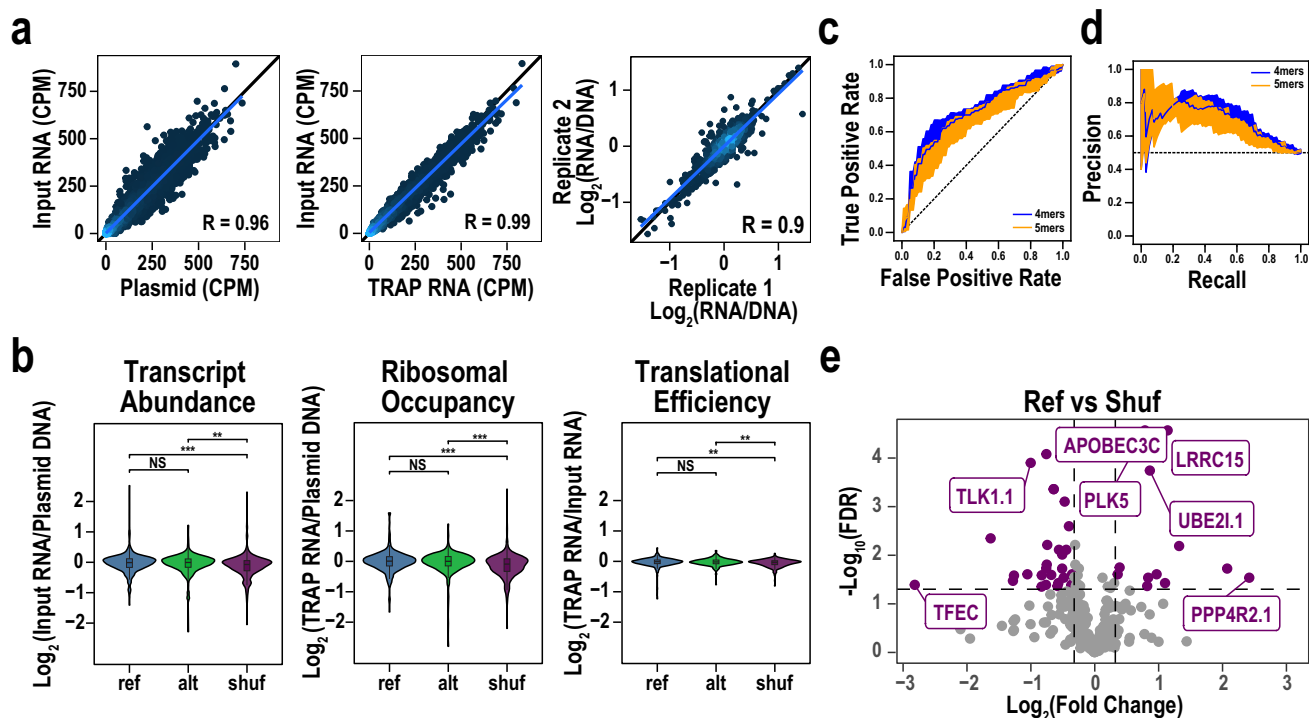
**Fig. 1 3'UTR library design and delivery.** **a** MPRA library constructs were designed with a CMV promoter (prom) driving the TdTomato reporter, followed by the 3'UTR oligo reference or alternative sequence (with or without variant, respectively) that is uniquely barcoded (BC). **b** All elements 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 components 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.

all experiments in paper are summarized in Supplementary Data 2). Correlations of either RNA measure with barcode abundance in recovered plasmid libraries averaged 0.99 (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, and inter-replicate correlation of this value was also high (0.9 PCC). This revealed variation in steady state RNA abundance across elements, with 99% of elements spanning  $-1.76$  to  $1.14$   $\log_2$ -normalized expression (RNA/DNA) [Fig. 2b], indicating that the sampled UTR elements exhibit a 7-fold range in transcript abundance as measured by our assay.

Normalizing TRAP RNA abundance by DNA copy number revealed a similar dynamic range 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 created a proxy for Translation Efficiency (TE), defined here as  $\log_2$  TRAP/Input counts. This showed a narrow dynamic range from  $-0.46$  to  $0.29$ , indicating 3'UTR element effects on translation regulation are more subtle than on transcript abundance. An interesting observation arose from a pairwise comparison of genome-derived reference elements to

GC-matched shuffled control sequences. Specifically, random sequences had both lower transcript abundance (Wilcoxon signed-rank  $p = 1.69 \times 10^{-4}$ ) and TE ( $p = 1.67 \times 10^{-4}$ ) than their corresponding reference sequences [Fig. 2b]. This suggests that genomic sequences generally promote higher steady-state transcript abundance and ribosome occupancy 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, and the effect of the alternate allele might be in either direction.

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 elements that increase or decrease RNA levels, we trained k-mer support vector machines (SVMs)<sup>25</sup> to classify the 200 highest-expressing elements from the 200 lowest-expressing elements, pooling Ref and shuffled (Shuf) sequences. 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



**Fig. 2** Screen in mouse neuroblastoma cell line identifies variants that alter steady state transcript abundance. **a** Scatter plots showing correlation between replicates of Input RNA vs plasmid DNA, Input RNA vs TRAP RNA, and Expression in Biological Replicate 1 vs. Replicate 2 ( $\text{Log}_2$  Input RNA/DNA). **b** Pairwise comparison of expression value distribution among Ref, Alt, and Shuf sequences in Transcript Abundance, Ribosomal Occupancy, and Translation Efficiency data sets.  $**p < 0.01$ ,  $***p < 0.001$  for Wilcoxon signed-rank test. Boxes on boxplots represent first quartile, median, and third quartile, whiskers represent minimum and maximum ( $Q1 - 1.5 \times$  interquartile range and  $Q3 + 1.5 \times$  interquartile range, respectively). **c** Receiver/operator and **(d)** precision recall 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  $\text{log}_2$  FC (x-axis). Horizontal dashed line corresponds to FDR 0.05 and vertical dashed lines correspond to  $\text{log}_2$  FC equivalent to 25% change in expression. Figure based on  $n = 6$  replicates. Full results list can be found in Supplementary Data 3, worksheet 1, and QC in Supplementary Data 2.

characteristic (AUROC) of between 0.707 and 0.663, for 4-mer and 5-mer models respectively, and an area under the precision recall curve (AUPRC) between 0.693 and 0.669 for these same models [Fig. 2c, d]. Models fit on random labels could not classify the data (AUROC between 0.486 and 0.521) [Supplementary 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<sup>26</sup> 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. It also indicates that for the effect sizes studied here, any effects of individual barcodes on transcript abundance were sufficiently small to not obscure the consistent biology detected by the SVM.

To determine which sequence motifs might be driving differences in RNA levels we screened for motifs enriched in the highest to the lowest 10% of the elements. We identified a variety of similar U/A rich motifs significantly enriched in the lowest expressed elements [Supplementary Data 3, Sheet 1], which are predicted binding targets of known negative regulators of RNA stability and expression such as TIA1<sup>27–30</sup>, QKI<sup>31</sup>, and PCBP2<sup>32,33</sup>. Likewise, we also identified two motifs found in more highly expressed elements, CCUUUCC and CCAACCC, predicted to be bound by PCBP1 and HNRNPK respectively, or

other RBPs with affinity to these motifs. While many RBPs share similar binding sequences it is difficult to ascertain which might be driving the effects seen here, these findings are consistent with the effects measured here being driven by sequence features of the individual elements.

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 54 were significantly different (Benjamini-Hochberg FDR < 0.05) with a median 1.65-fold change in expression [Fig. 2e, Supplementary Data 4]. Thus, genomic sequences produce a specific level of activity upon which allelic effects are expected to act. Of the 303 tested comparisons, 40 showed both a significant difference and a > 25% magnitude change in expression. Of the significant changes, 38 were downregulating. Assuming equal probability of up- and down-regulation, this is more than expected by chance (hypergeometric  $p = 0.0033$ , OR = 1.69), again reflecting the relative greater propensity for genomic-derived UTR tiles to enhance steady-state reporter expression. Contrasting effects on transcript abundance with ribosome occupancy again revealed that variant effects on TE tended to be much smaller.

Finally, we sought to determine which of these changes in RNA level might also alter protein production. We therefore cloned 28 of these reference and variant sequences into luciferase reporters and assessed protein production via transient transfection in N2a cells. Across the constructs, we observed a Pearson correlation coefficient of  $R = 0.65$ ,  $p = 0.0002$  between our MPRA expression and protein production, indicating most of the effects we observe

at the RNA level have a similar consequence on final protein levels [Supplementary Fig. 3]. Overall, our cell line assay confirmed reproducibility and robustness of our measurements of element level activity by our 3'UTR MPRA design, motivating applying the approach to specific cell types in vivo.

**Cre-dependent MPRA reproducibly measures functional effects of several hundred UTR elements 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<sup>34</sup> that with AAV9 delivery we get widespread viral transduction in the neocortex and mainly target neurons and astrocytes. 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 [Supplementary Figure 4]. Thus, as packaging had no adverse effects on the composition of the library we moved forward with delivery in vivo.

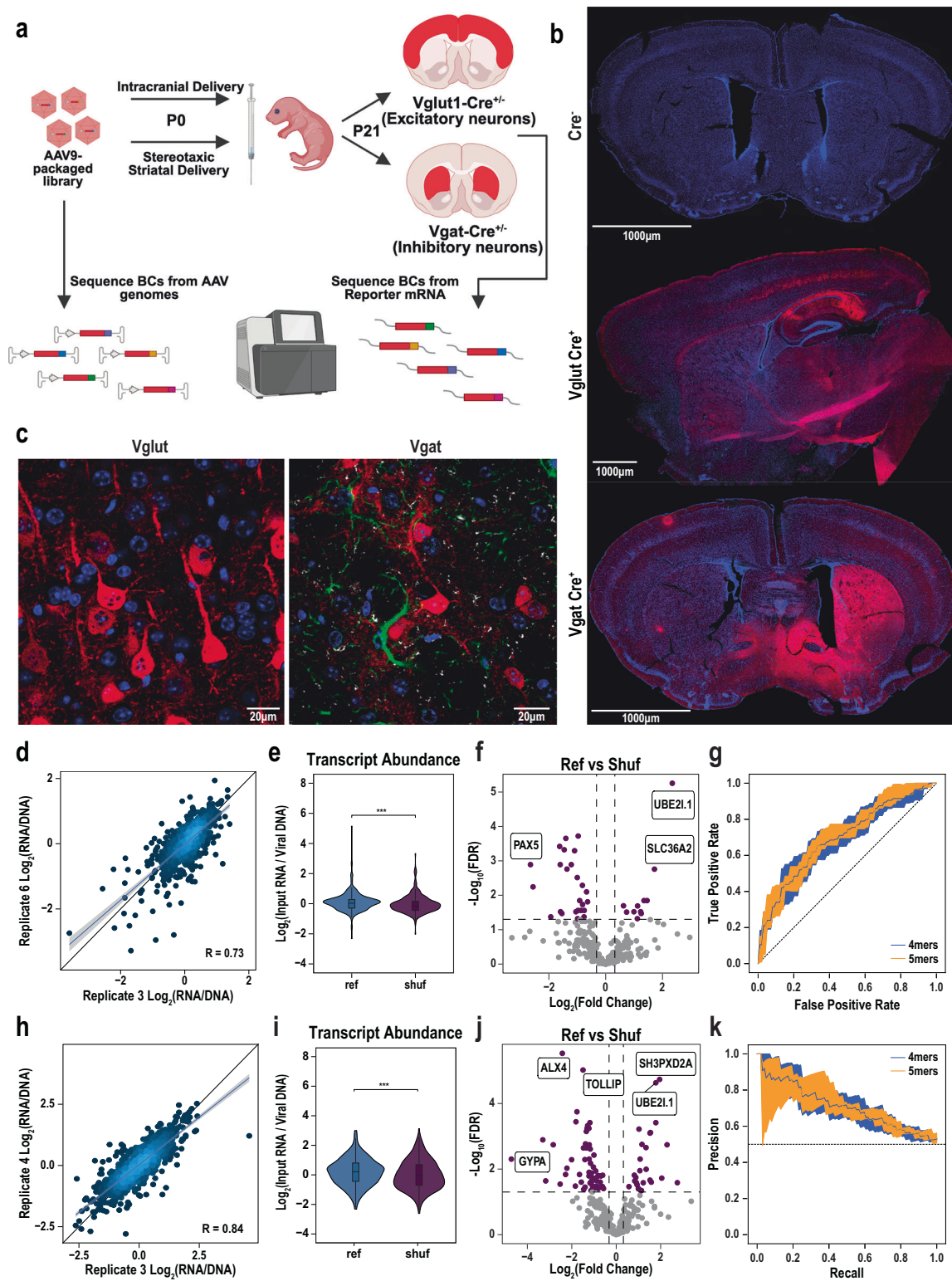
Bioinformatic analysis of the expression patterns of genes associated with autism have revealed a correlation structure of two loose modules - a module enriched for chromatin regulators with peak expression in immature excitatory neurons, and a module of synaptic-related proteins, with peak expression during critical periods of postnatal synaptogenesis and pruning<sup>7,35–37</sup>. Therefore, we first attempted to deliver the library to two neuronal subtypes, layer V pyramidal neurons and cortical GABAergic interneurons, during this pruning period by using RBP4 and VGAT Cre driver lines, respectively. However, especially when using an untargeted P1 injection strategy, we discovered that only a small fraction of the delivered elements were recovered, the 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 ( $PCC < 0.2$ ). This could either be due to high biological variability or a technical effect termed 'jackpotting' which here we define as having barcodes whose final measurement in the MPRA sequencing library does not reflect their starting abundance in the sample RNA, likely because only a subset of the barcodes were sampled at a particular step in the library preparation. We conducted extensive experiments to determine at which step such jackpotting might occur (see methods), and traced it to the very first step. This indicated that having the Cre recombinase only in a small fraction of the cells resulted in a very low 'library density' in the total RNA, meaning only a very small fraction of the RNA molecules in the sample came from the MPRA library.

Since RBP4-positive and VGAT-positive cells made up a small population of cells in the mouse brain, to increase the library density we delivered the AAV library to a well-characterized excitatory neuron specific Cre line (*Vglut1-IRES2-Cre-D*<sup>38</sup>; *Vglut1*) at P0-P2 [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 [Fig. 3b, c]. While we did not confirm the expression of the MPRA library here with marker colabeling, we note that recent single-cell atlasing of this same mouse line identified that over 99% of cells with Cre activity in cortex cluster as excitatory pyramidal neurons<sup>39</sup>. Importantly, Cre negative littermates showed no expression of the library, confirming cell-type specificity [Fig. 3b]. Next, an additional 11 animals' (6 males and 5 females) cortices were collected for RNA at P21. We sequenced, in all, 11 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 expression between biological replicates on average exceeded 0.70 (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 also motivated our doubling of the number of replicates in vivo relative to in vitro). 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 313 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  $p = 2.99 \times 10^{-4}$ ) than their corresponding Ref sequences, as observed in N2as [Fig. 3e]. Of the 301 testable Ref-Shuf comparisons, 36 showed a significant difference in expression. While we observed a 25:11 ratio of downregulation:upregulation effect of the reference sequences vs. shuffled sequences, the enrichment was not significant (hypergeometric test,  $p$ -value = 0.14, OR = 1.19). [Fig. 3f, Supplementary Data 4] Finally, we again used k-mer SVMs to determine if there were sequence features that predicted in vivo activity and achieved an AUROC between 0.696 and 0.706, for 4-mer and 5-mer models respectively [Fig. 3g], and an AUPRC between 0.708 and 0.704 for these same models [Fig. 3k], comparable to the SVMs trained on in vitro activity. Thus, for the large effect sizes observed when changing all sequences in an element, our approach was sufficiently powered.

**In vivo Cre-dependent MPRA works reproducibly in more than one cell type.** We next sought to determine if the method was effective across cell types. We returned to the VGAT (GABAergic) mouse line but with a modifications of our prior approach - we focused our delivery and dissection on the striatum, an area that is both of high interest for autism<sup>40</sup>, but also where >90% of the neurons are GABAergic medium spiny neurons, making it easier to deliver the library into a large fraction of the of cells. This resulted in targeted striatal expression [Fig. 3b] in cells with the morphology of medium spiny neurons, which was absent from astrocytes and oligodendrocytes (GFAP and CNPase staining, respectively) [Fig. 3c]. We saw better correlations than prior experiments [Fig. 3h], and again genomic sequences generally had greater abundance than shuffled controls [Fig. 3i], with dozens having significant activity [Fig. 3j, Supplementary Data 4]. In all, this demonstrates that the Cre-dependent MPRA is effective across more than one cell type, and that we were well-powered to detect the effect sizes seen when comparing reference sequences to random sequence controls.

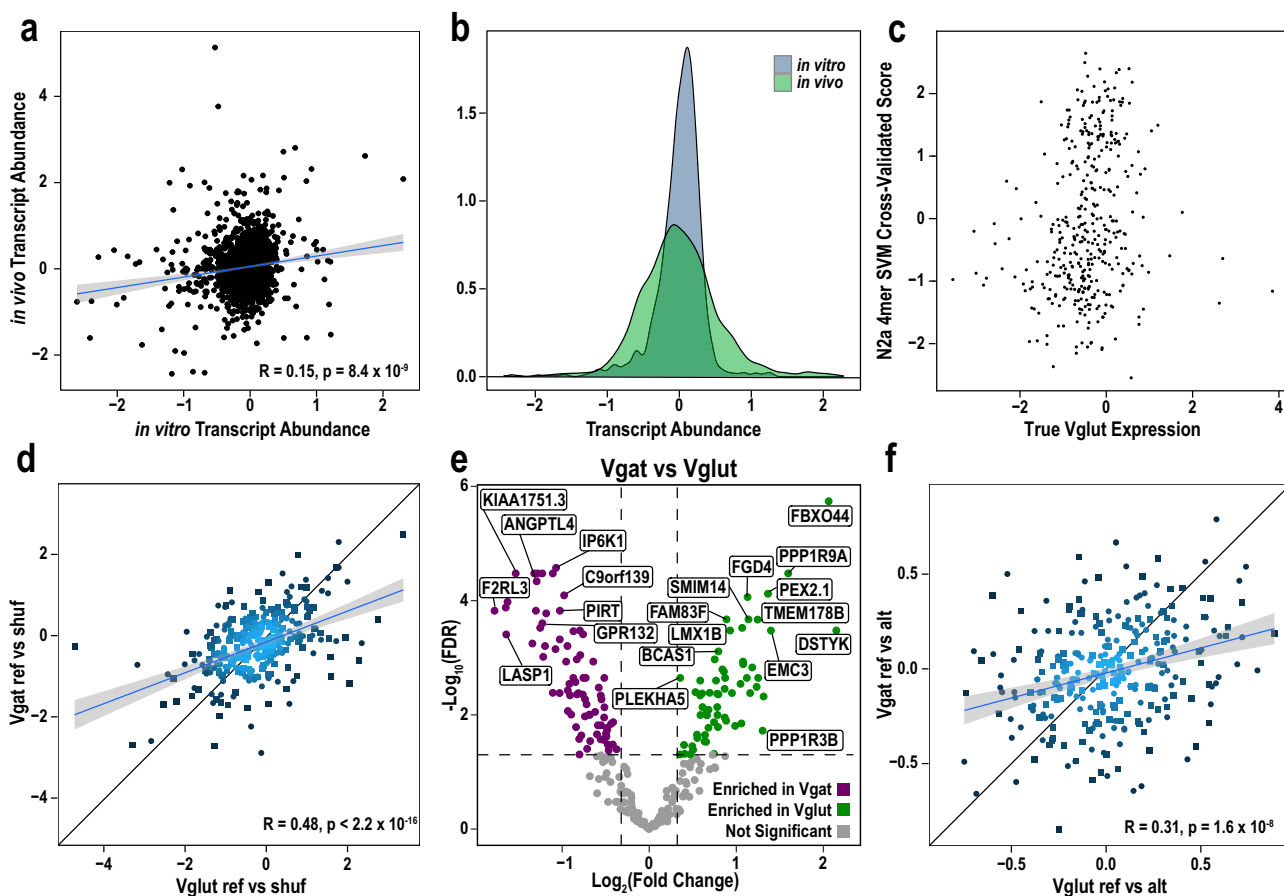
Thus, the Cre-dependent MPRA should allow quantification of the impact on transcript abundance of a given element in specific cell types in the brain. This is essential because neurons have vastly different expression of *trans*-acting factors (e.g., TFs, RBPs, miRNAs) than cell lines. This is highlighted by the low correlation of expression values of element activity across N2As compared to pyramidal neurons [Fig. 4a]. Furthermore, transcript abundance spanned a broader range in the pyramidal neurons (Brown-Forsythe Levene-type  $p < 2.2 \times 10^{-16}$ ), highlighting the possibility that a more complex regulatory environment may contribute to a greater dynamic range [Fig. 4b]. Finally, the cross-validated SVM scores of the N2a activity are uncorrelated to the observed activity in pyramidal neurons [Fig. 4c], consistent with cell type-specific factors regulating UTR activity through interaction with specific sequences. This highlights the need to assess the function of noncoding elements in



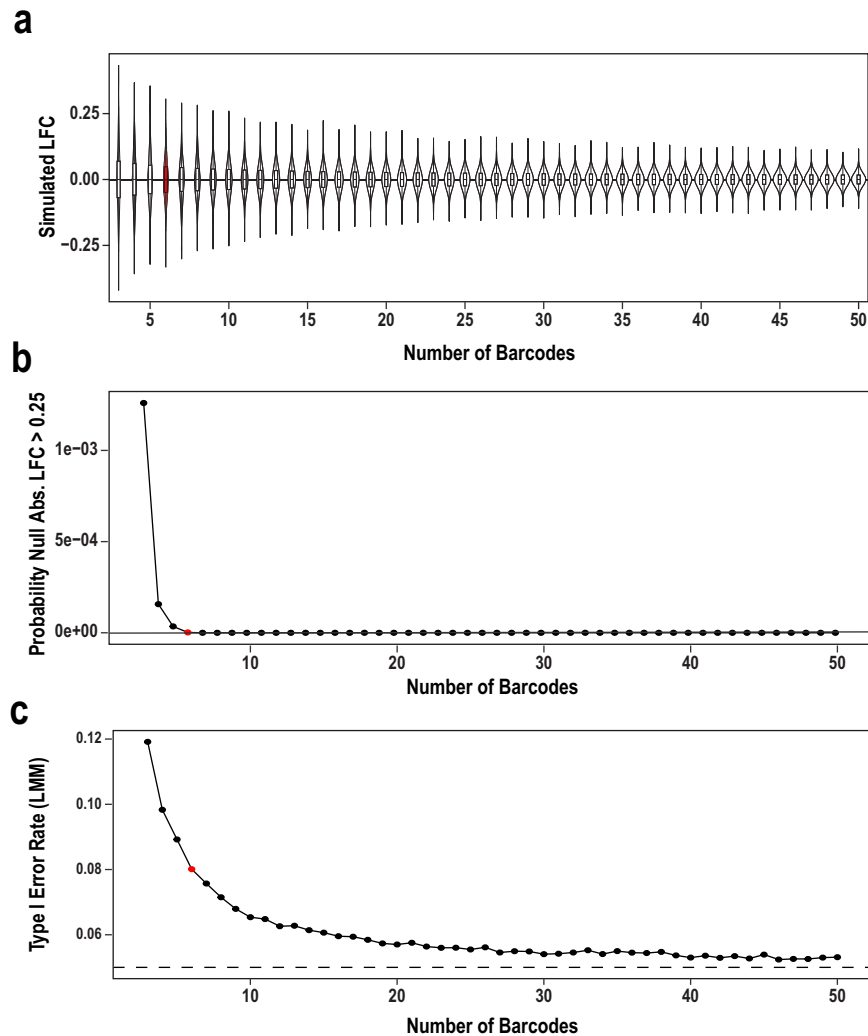
multiple contexts, and especially focusing on contexts where noncoding variants for a specific disease are most likely to act. Comparing our two neuronal types together reveals a somewhat better correlation, consistent with a somewhat shared regulatory milieu [Fig. 4d], but with some differences in element activity detectable between them [Fig. 4e]. As above, to determine which sequence motifs might be driving differences in RNA levels between medium spiny neurons and pyramidal neurons, we screened for motifs enriched in the highest to the lowest 10%

the elements in this comparison. We found a notable enrichment in U rich sequences in those elements expressed higher in the medium spiny neurons, suggesting these neurons may have lower activity of RBPs that bind to such sequences and destabilize RNA, such as TIA, or higher level of a stabilizing protein such as HUR<sup>41,42</sup> [Supplementary Data 3]. Finally we examined the effect sizes and correlations of potential allelic effects, and noted they were somewhat less correlated, perhaps because they were generally smaller and more variable than element effects [Fig. 4f].

**Fig. 3 Screen in excitatory neurons in the mouse brain identifies elements that alter steady state transcript abundance.** **a** MPRA library was packaged into AAV9 and delivered into perinatal mouse cortices via intracranial injection or bilaterally into striatum via stereotaxic injection and later harvested at P21 for RNA extraction. Libraries were prepared from AAV genomes and reporter mRNA, and barcodes (BC) were counted. Panel was created with Biorender.com. **b** Immunofluorescence of P21 brain showing localization of tdTomato expression (from MPRA library) in Cre lines and wildtype control (Cre-). Nuclei counterstained with DAPI(blue). Scale bars represent a distance of 1000  $\mu$ m. **c** Immunofluorescence demonstrating expression of MPRA library in morphological pyramidal neurons (Vglut) and medium spiny neurons (Vgat). There is no overlap with markers of astrocytes (GFAP) or oligodendrocytes (CNPase). Scale bars represent a distance of 20  $\mu$ m. **d** Scatter plot showing correlation between replicates for mean-normalized expression in excitatory pyramidal neurons (Log<sub>2</sub> RNA/DNA, barcode averaged). Shaded region indicates 95% confidence interval around linear regression line. **e** Pairwise comparison of Ref, and Shuf sequence expression in excitatory neurons. \*\*\* $p < 0.001$  for Wilcoxon signed-rank test. Boxes on boxplots represent first quartile, median, and third quartile, whiskers represent minimum and maximum (Q1 - 1.5 \* interquartile range and Q3 + 1.5 \* interquartile range, respectively). **f** Volcano plot for Ref vs Shuf element expression in excitatory neurons in library showing significance (y-axis) vs log<sub>2</sub> FC (x-axis). Horizontal dashed line corresponds to FDR 0.05 and vertical dashed lines correspond to log<sub>2</sub> FC equivalent to 25% change in expression. Full results list can be found in Supplementary Data 4 and 5, worksheet 3. **g** Receiver/operator (upper panel) and precision/recall (lower panel) curves for k-mer SVMs to classify high and low expressing elements in excitatory neurons. Shaded area represents 1 standard deviation based on five-fold cross-validation. **h** Scatter plot showing correlation between replicates for mean-normalized expression in inhibitory neurons (Log<sub>2</sub> RNA/DNA, barcode averaged). Shaded region indicates 95% confidence interval around linear regression line. **i** Pairwise comparison of Ref, and Shuf sequence expression in medium spiny neurons, \*\*\* $p < 0.001$  for Wilcoxon signed-rank test. Boxes on boxplots represent first quartile, median, and third quartile, whiskers represent minimum and maximum (Q1 - 1.5 \* interquartile range and Q3 + 1.5 \* interquartile range, respectively). **j** Volcano plot for Ref vs Shuf element expression in medium spiny neurons. Full results list can be found in Supplementary Data 4 and 5, worksheet 4.  $n = 6$  Vgat, 11 Vglut animals. **k** Precision/recall curves for k-mer SVMs to classify high and low expressing elements in excitatory neurons.



**Fig. 4 Cell type specific influences on regulatory element expression.** **a** Scatter plot of *in vitro* N2a vs. *in vivo* pyramidal neuron (Vglut) expression values show poor correlation indicate importance of cell type. **b** Comparison of *in vitro* and *in vivo* expression distributions, Brown-Forsythe Levene-type test for difference in variance  $p < 2.2 \times 10^{-16}$ . **c** N2a SVM score vs. true Vglut expression show lack of correlation - indicating *in vitro* data cannot predict *in vivo* values. **d** Scatterplot comparing pyramidal neurons and medium spiny neuron (Vgat) expression values. **e** Volcano plot identifies elements that have differential activity between pyramidal and medium spiny neurons. **f** Scatterplot shows weaker correlation and smaller effect sizes for allelic effects when compared to element effects (**d**). Case and control variants are labeled as circles and squares, respectively. For all statistical reporting,  $n = 6$  for Vgat animals and  $n = 11$  for Vglut animals. Shaded regions on (**a**, **d**, **f**) indicates 95% confidence interval around linear regression line.



**Fig. 5 Simulated tests sampling control data to model barcode effects.** Simulations drawing sets of barcodes for ‘allelic’ comparison using a barcode count of three through fifty for each allele (1,000,000 iterations, drawing from ~100 Barcodes tagging the same control element from a previous study of ours<sup>43</sup>). Red data points indicate six barcodes. **a** Violin plots showing the range and frequency of log<sub>2</sub> fold-changes for randomly sampled data for the range of barcodes. Boxes on boxplots represent first quartile, median, and third quartile, whiskers represent minimum and maximum (Q1 - 1.5 \* interquartile range and Q3 + 1.5 \* interquartile range, respectively). **b** Plot detailing the probability of obtaining a log<sub>2</sub> fold change less than or greater than 0.25 for randomly samples data for the range of barcodes. **c** Plot detailing the Type I Error Rate of a linear mixed model for randomly sampled data for the range of barcode numbers.

**Comparison of effect sizes for barcodes relative to alleles.** We next wanted to determine if our method was able to reliably detect smaller effect sizes that might be associated with single nucleotide variants, which comprise much of our variant set. One concern is barcodes - while the internal replication they provide can improve statistical power, it is possible that the effects of changing a 8-9 nucleotide barcode on transcript abundance may be more substantial than the 1 bp variant it tags. Including multiple barcodes per replicate is traditionally thought to allow correction for this by averaging or regression. Therefore, we modeled using random sets of barcodes drawn for the controls in a recent publication<sup>43</sup> to determine rates of false-positive ‘allelic effects’ driven by barcodes only [Fig. 5]. We found that linear mixed modeling with 6 barcodes would be sufficient to remove the majority of barcode effects for most experiment designs, leaving a false positive rate of 8% [Fig. 5c], and generally small effects [Fig. 5a, b]. This could be tolerable for many designs, however, if the design of the experiment is a screen for extremely rare functional effects (i.e., such as the rare autism variants

examined here), this 8% false positive rate produced by 6 barcodes may be too high. Indeed, this is similar to the number of hits seen when testing for allelic effects here [e.g., 2–6 per assay, Supplementary Data 5]. Thus we do not confidently report on the allelic effects from this experiment. And while 1000’s of barcodes per variant can be added by PCR<sup>18,44,45</sup>, the corresponding increase in library complexity would lead to substantial jackpotting in the cell-type specific in vivo context, where delivering each barcode to enough cells for robust measurement is a challenge. Thus, 6 barcodes might be sufficient for examining the large effects seen when measuring large effect size changes (e.g., the differences between elements) but more barcodes would be needed to assess smaller allelic effects, or when screening rare variants where only a small fraction are expected to be functional.

## Discussion

Here we describe the development of a cell-type specific in vivo MPRA. We demonstrate that the method is sensitive enough to



identify reproducible effects for hundreds of elements in parallel. It should 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<sup>18</sup> and other perturbations<sup>46</sup>. It, with additional barcoding or other solutions, can also be adapted to examine allelic effects as well. Thus, the approach should have both translational and basic science applications. Furthermore, while the current library used limited UTR lengths (120 bp, due to synthesis limits at the time), the synthesis-length of oligonucleotide arrays continues to increase and new methods for assembling these into even larger fragments have recently proven viable<sup>47</sup>. Thus, it should be routinely applicable to even larger elements soon.

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 the 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, here, we targeted the striatum, where a greater concentration of GABAergic cells exist), 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. Indeed, we did recently demonstrate the ability to measure allelic effects using TRAP in an enhancer context, especially when the library was delivered in a concentrated manner to a single anatomical structure so that library density is higher<sup>48</sup>. 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 elements in the regulatory context of mature cortical and striatal neurons, essential cell types for many CNS diseases. One key factor in the design, however, is the number of barcodes, as there is a clear tradeoff between adding more barcodes and elements which gains in experimental efficiency per animal, yet decreases reproducibility due to jackpotting, or barcode effects. For screening libraries where only a small fraction are expected to be functional, it may be preferred to reduce library complexity and barcode effects by removing barcodes entirely (letting the UTR serve as the barcode), or for designs that require barcodes (e.g., assessing enhancer variants) aim for ~20 barcodes per allele [Fig. 5]. Of course, libraries with larger effect sizes than SNPs (e.g., varying entire elements) do not need such a high level of stringency.

With regards to the role of these specific alleles in autism, we are hesitant to make any specific claims since the number of variants showing significant effect was similar to what would be expected due to barcode effects in our simulation studies. Thus, we can't be certain which allelic effects to confidently report. Nonetheless, there are two clear conclusions we can make. First, we can rule out large numbers of high-effect size UTR variants as a common cause for autism - if they were a common cause, we would have seen an excess of functional variants coming from the proband's alleles relative to the unaffected sibling alleles, and we did not. Note - this does not rule out that in some rare instances a UTR mutation could be causal, but it will not be a frequent occurrence. Second, we can report that the typical effect sizes of single base mutations is small - even the significant 'allelic' effects detected were typically less than 25% changes, while differences in expression due to elements ranged over 8 fold. Thus, alleles with large functional effects will be rare.

Regardless, for elements of large effects, these cell type-specific MPRA should enable in vivo identification of functional elements across Cre-defined cell types. This is important because the function of non-coding elements is strongly impacted by the suite of transcriptional regulators expressed in each cell type. Thus, this method presents the opportunity 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.

## Methods

**Animal models.** All procedures involving animals were approved by the Institutional Animal Care and Use Committee (IACUC) at Washington University in St. Louis, MO. Veterinary care and housing was 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<sup>49</sup>), *Slc32a1tm2(cre)Lowl/J* (catalog #16962, The Jackson Laboratory; RRID:IMSR\_JAX:016962<sup>50</sup>), 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 (0.5 M Tris-HCl pH 8.8, 0.25 M EDTA, 0.5% Tween-20, 4 uL/mL of 600 U/mL Proteinase K enzyme) for 1 h to overnight. This is followed by heat denaturation at 99 C for 10 min. 1 uL of the resulting lysate was used as a template for PCR with with 500 nM forward and reverse primers, using 1x Quickload Taq Mastermix (NEB) with the following cycling conditions: 94 1 min, (94 30 s, 60 30 s, 68 30 s) 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.

In order to determine if the elements in our library came from genes expressed in *Vglut* and *Vgat* regions specifically in excitatory neurons and medium spiny neurons we examined two single cell datasets<sup>51,52</sup> GSE171977, and GSM4471659. The

mean expression for each gene across all cells was calculated in both datasets, and ‘expressed’ was defined as exceeding the median of this value. The data was then subsetted to only excitatory neurons or medium spiny neurons and the mean expression for each gene was calculated and expression was determined. Genes were then converted back into their mouse orthologs using *ensembl* and these true false values were paired with their respective elements on Supplementary Data 4 and 5.

The specific oligo sequences with barcodes designed for this library are provided in [Supplementary Data 6]. 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 DH5 $\alpha$  (NEB). Outgrowths were plated on LB agar plates with 100  $\mu$ g/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 h in LB with carbenicillin before preparing glycerol stocks and maxi preps (Qiagen).

**Cell culture.** Mouse neuroblastoma N2a cells (ATCC) were maintained at 5% CO<sub>2</sub>, 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 h 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% confluent by the time of transfection. As this is a Cre-inducible library, 23  $\mu$ g of total plasmid was transfected, consisting of equimolar ratios of the MPRA library, an EF1a-DIO-EGFP-RP110a construct, and an Efla-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 min prior to transfection. Transfection medium was replaced 12 h

following transfection, and cells were harvested for TRAP after an additional 24–36 h.

TRAP was performed as described in ref. <sup>24</sup> with minimal modification. Briefly, cells were incubated in 100  $\mu$ g/mL cycloheximide (Sigma) for 15 min at 37 °C prior to harvest. Cells were rinsed twice with 5 mL of DMEM 100  $\mu$ g/mL cycloheximide before being lifted into 5 mL of DMEM 100  $\mu$ g/mL cycloheximide. Cells were pelleted by spinning at 500 x g for 5 min 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 MgCl<sub>2</sub>, 0.5 mM dithiothreitol, 100  $\mu$ g/ml CHX, protease inhibitors, and RNase inhibitors) and cells were lysed on ice. Lysates were clarified by centrifugation at 2000 x g for 10 min at 4 °C. DHPC (Avanti) was added to a final concentration of 30 mM, and lysates were incubated on ice for 5 min. 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<sup>53</sup>. The beads were incubated for 4 h 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 MgCl<sub>2</sub>, 0.5 mM dithiothreitol, 100  $\mu$ g/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.

**AAV9 vector production.** The packaging cell line, HEK293, was maintained in Dulbecco’s modified Eagles medium (DMEM), supplemented with 5% fetal bovine serum (FBS), 100 units/ml penicillin, 100 mg/ml streptomycin in 37 °C incubator with 5% CO<sub>2</sub>. The cells were plated at 30–40% confluence in CellSTACS (Corning, Tewksbury, MA) 24 h before transfection (70–80% confluence when transfection). 960  $\mu$ g total DNA (286  $\mu$ g of pAAV2/9, 448  $\mu$ g of pHelper, 226  $\mu$ g of AAV transfer plasmid) were transfected into HEK293 cells using polyethylenimine (PEI)-based method<sup>54</sup>. The cells were incubated at 37 °C for 3 days before harvesting. The cells were then lysed by three freeze/thaw cycles. The cell lysate was treated with 25 U/ml of Benzonase at 37 °C for 30 min and then purified by iodixanol gradient centrifugation. The eluate was washed 3 times with PBS containing 5% Sorbitol and concentrated with Vivaspin 20 100 K concentrator (Sartorius Stedim, Bohemia, NY). Vector titer was determined by qPCR with primers and labeled probe targeting the ITR sequence<sup>55</sup>. Titers used here ranged from 1 to 5  $\times$  10<sup>13</sup> vg/ml.

**in vivo MPRA.** For excitatory neurons, 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 min. An aliquot of the MPRA library packaged in AAV9 (~10<sup>12</sup> vg/ $\mu$ L) was drawn up in a 33 G Hamilton syringe with a 1 mm needle. Pups were

brought up to the needle and 1  $\mu$ L 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 min). After recovery, pups were placed back into the cage with the mother and monitored every 24 h for one week. At P21 brains were harvested and cortex was dissected away from the rest of the brain for extraction of RNA.

For striatal medium spiny neurons *Slc32a1tm2(cre)Lowl/J* litters were subjected to intracranial stereotaxic striatal injections for delivery of the library packaged in AAV9. P0-P2 pups were incubated on ice to anesthetize by inducing hypothermia for ~10 min, after which they were placed in a 3D printed adaptor, as per Olivette et al.<sup>56</sup>. An aliquot of the MPRA library packaged in AAV9 (~10<sup>12</sup> vg/ $\mu$ L) was drawn up in a 33 G Hamilton syringe with a 1 mm needle. 1  $\mu$ L of virus was injected at one position per brain hemisphere hemisphere (2 total injections per pup) using coordinates  $x = \pm 1.0$  mm,  $y = + 0.7$  mm,  $z = +1.5$  mm at a rate of 0.5  $\mu$ L/minute, with 1 min dwell time. Pups were taken directly to the warming pad until fully recovered (~20 min). After recovery, pups were placed back into the cage with the mother and monitored every 24 h for one week. At P21 brains were harvested and striatum was dissected away from the rest of the brain for extraction of RNA.

For our initial pilot studies in *RBP4* and cortical inhibitory neurons, injections were as conducted above for *Vglut1-IRES2-Cre-D*, but these experiments showed too low of correlation between replicates to proceed to further analysis due to jackpotting. 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 [Supplementary Fig. 5A], and compared this to the low reproducibility observed between biological replicates [Supplementary Fig. 5B]. Taking a single RNA sample and doing two separate cDNA synthesis reactions for independent sequencing libraries resulted in jackpotting (PCC < 0.4) [Supplementary Fig. 5C]. Taking cDNA from a single sample and amplifying it in two independent reactions for library preparation also led to jackpotting samples (PCC < 0.4) [Supplementary Fig. 5D]. 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) [Supplementary Fig. 5E]. This was the case for reactions split at the final indexing PCR as well (PCC > 0.9) [Supplementary Fig. 5F]. 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 and unique molecular identifiers (UMIs) at the cDNA synthesis step in order to precisely quantify and eliminate PCR duplicates. Resulting technical replicates were evaluated to determine if they suppressed jackpotting results and improved reproducibility as measured by PCC. However, after sequencing the resulting libraries and computationally collapsing UMIs, we found that this did not fundamentally improve jackpotting (best PCC < 0.34). 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 pipetted (100 ng) from our total RNA from the brain, relatively few barcode molecules were present. Consistent with this, increasing input RNA up to 1  $\mu$ g reduced jackpotting effects, but still resulted in relatively low sample correlations (PCC < 0.4). The possible solutions are to 1) increase the amount of RNA going into cDNA synthesis, which improved

things to point (from PCC < 0.2 to PCC < 0.4) or 2) increase the library density by getting the library into a higher fraction of cells contributing to the total RNA, which is the strategy we adopted for the subsequent experiments.

**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 (GCATAAAA AACAGACTACATAATACTG) 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 (GCATGGACGAGCTG-TACAAG) 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 h at 37 °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 2  $\times$  150 next generation sequencing on an Illumina NovaSeq.

**BC counting and normalization.** Sequencing reads were trimmed using cutadapt v1.16<sup>57</sup> and aligned to the library reference sequences using bowtie2 v2.3.5<sup>58</sup> 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<sup>59</sup> as counts per million (CPM) prior to downstream analysis.

Measures of transcript abundance were calculated as the log<sub>2</sub>-transformation of the ratio of each barcode's abundance, in CPM, from Input RNA over DNA, within each replicate. Barcodes with fewer than ten counts in either the RNA or DNA library were excluded from analysis. Similarly, ribosomal occupancy and translation efficiency were calculated by normalizing TRAP RNA to DNA counts and TRAP RNA to Input RNA counts, respectively. To calculate an element-wise measurement of transcript abundance or translation activity, a linear mixed effect model was fit which accounted for outlier barcode effects. Barcode-level measurements from each replicate were fitted to the formula Activity ~ (1 | BC) using the lmer package in R where Activity may be either transcript abundance, ribosomal occupancy, or translation efficiency. The model intercepts were taken for each element as the summarised measure of activity.

**Element filtering and differential activity analysis.** Differential effects of allele on transcript abundance and translation efficiency were each tested using a linear mixed effects model fitting random intercepts for barcode. This model was implemented using the lmer package in R with the formula Activity ~ Allele + (1 | BC), where Activity may be either transcript abundance or translation efficiency. P-values were computed using a likelihood ratio test (LRT) with Activity ~ (1 | BC) as the reduced model, and corrected for multiple comparisons by using the p.adjust function in R to apply the Benjamini-Hochberg procedure for false discovery rate.

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 set 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. An arbitrary count threshold minimum of 10 counts was enforced. 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 at least four replicates. For the in vivo MPRA, variants must be present with both alleles having five barcodes with at least 10 counts from both RNA and DNA in six replicates.

**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) or rabbit anti-RFP and goat anti-GFAP (1:500; ABCAM catalog #ab53554) and mouse IgG1 anti-CNPase (1:500; Millipore-Sigma catalog# MAB 326 R) 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) or donkey anti-rabbit Alexa Fluor 568, donkey anti-goat Alexa Fluor 488 (1:1000, Jackson Immunoresearch catalog # 705-546-147), goat anti-mouse IgG1 Alexa Fluor 647 (Invitrogen catalog # A21240), 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<sup>60</sup>.

**Machine learning.** Gapped k-mer SVM models were fit using gkmSVM<sup>25</sup> 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<sup>61</sup>.

**Luciferase validation experiments.** 3'UTR elements were cloned into a custom-designed dual luciferase reporter plasmid that contains derivatives of Promega Nano-Luc and Firefly plasmids. Briefly, the 3'UTR elements were cloned 3' of a Nano-Luciferase CDS. The Nano-luciferase CDS bears a PEST degradation signal and contains a chimeric intron<sup>62</sup>, and is under the transcriptional control of the CMV immediate early enhancer and promoter, while Firefly luciferase is under control of the human PGK promoter. The Nano-Glo Dual Luciferase Reporter Assay (Promega) was used to assess first Firefly levels then Nano-Luciferase levels. Nano-Luciferase levels were then normalized to Firefly levels to achieve a Nano-Luciferase activity value relative to the amount of transfected plasmid. Control plasmid not bearing any luciferase sequences was separately transfected to control for plate background luminescence. Reference and variant elements were assessed for statistically different activities using mixed effect modeling from the lme4 R package, with the plate ID being used as a random effect to account for batch effect between

biological replicates, and activity values were compared to a blank construct bearing no inserted 3'UTR.

**MEME suite motif search and comparison.** Motifs in high- or low-expression elements, or reference elements with significant higher or lower expression in excitatory neurons, were discovered with the MEME suite of tools. Briefly, for high vs. low expression elements, the list was split into high and low, and 10% of the topmost with regard to effect (most positive for positive, most negative for negative) were given to MEME, using the opposite list as the negative control. The differential enrichment option was selected with -de, and the -norand option was selected to keep the list ordered from greatest effect to least effect in the lists, and derivatives of the following command were run in a local installation: `./meme high.txt -neg low.txt -rna -objfun de -norand -mod anr -minw 3 -maxw 50 -text -nmotifs 100 > results.txt`

For reference elements which were significantly higher expressed in excitatory vs. inhibitory neurons (and vice versa), a similar command was used to search for differential enrichment where the negative was the opposite list (all significantly higher in inhibitory were used as the negative for excitatory searching and vice versa). Outputs of discovered differentially enriched motifs were submitted to the Tomtom motif comparison tool on the meme-suite.org website, and queried for matches to human RNA binding protein. Reverse complements were not scored.

**Statistics and reproducibility.** Sample sizes, sequencing depth, etc. for all MPRA experiments are defined in Supplementary Data 2. Biological replicates were defined as individual wells (for cell culture experiments) or individual animals (for in vivo experiments), and sample sizes reported here reflect biological replicates. Technical replicates (independent preparations of the same biological materials) were only used where reported in figures to examine technical reproducibility of library preparation steps (e.g., Supplementary Fig. 5). Details of statistical tests used are provided in figure legends and methods sections above for each experiment type.

**Reporting summary.** Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

#### Data availability

MPRA libraries are available upon request. MPRA data are deposited with GEO at GSE186455. Source data underlying figures are provided in Supplementary Data 7.

#### Code availability

Code is available at Figshare: <https://doi.org/10.6084/m9.figshare.24093942><sup>63</sup>.

Received: 15 February 2023; Accepted: 18 October 2023;  
Published online: 13 November 2023

#### References

1. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022)
2. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
3. Matoba, N. et al. Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Transl. Psychiatry* **10**, 1–14 (2020).
4. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants Across Biological Contexts. *Biol. Psychiatry* **89**, 76–89 (2021).

5. Hevner, R. F., Hodge, R. D., Daza, R. A. M. & Englund, C. Transcription factors in glutamatergic neurogenesis: Conserved programs in neocortex, cerebellum, and adult hippocampus. *Neurosci. Res.* **55**, 223–233 (2006).
6. Pilaz, L.-J. & Silver, D. L. Post-transcriptional regulation in corticogenesis: how RNA-binding proteins help build the brain. *Wiley Interdiscip. Rev. RNA* **6**, 501–515 (2015).
7. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
8. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
9. Iossifov, I. et al. De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285–299 (2012).
10. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
11. O’Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
12. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
13. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
14. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
15. Turner, T. N. et al. Genomic patterns of De Novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
16. Mayr, C. Regulation by 3’-Untranslated Regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
17. Choi, J. et al. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* **11**, 2718 (2020).
18. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
19. Litterman, A. J. et al. A massively parallel 3’ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* <https://doi.org/10.1101/gr.242552.118> (2019).
20. Siegel, D. A., Le Tonqueze, O., Biton, A., Zaitlen, N. & Erle, D. J. Massively parallel analysis of human 3’ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3 Genes|Genomes|Genetics* **12**, jkab404 (2022).
21. Griesemer, D. et al. Genome-wide functional screen of 3’UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260.e19 (2021).
22. Schnütgen, F. et al. A directional strategy for monitoring Cre-mediated recombination at the cellular level in the mouse. *Nat. Biotechnol.* **21**, 562–565 (2003).
23. Heiman, M. et al. A translational profiling approach for the molecular characterization of CNS cell types. *Cell* **135**, 738–748 (2008).
24. Heiman, M., Kulicke, R., Fenster, R. J., Greengard, P. & Heintz, N. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nat. Protoc.* **9**, 1282–1291 (2014).
25. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using Gapped k-mer features. *PLOS Comput. Biol.* **10**, e1003711 (2014).
26. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
27. Piecyk, M. et al. TIA-1 is a translational silencer that selectively regulates the expression of TNF-alpha. *EMBO J.* **19**, 4154–4163 (2000).
28. Dixon, D. A. et al. Regulation of cyclooxygenase-2 expression by the translational silencer TIA-1. *J. Exp. Med.* **198**, 475–481 (2003).
29. Rodrigues, D. C. et al. MECP2 Is Post-transcriptionally Regulated during Human Neurodevelopment by Combinatorial Action of RNA-Binding Proteins and miRNAs. *Cell Rep.* **17**, 720–734 (2016).
30. Byres, L. P. et al. Identification of TIA1 mRNA targets during human neuronal development. *Mol. Biol. Rep.* **48**, 6349–6361 (2021).
31. Neumann, D. P., Goodall, G. J. & Gregory, P. A. The Quaking RNA-binding proteins as regulators of cell differentiation. *Wiley Interdiscip. Rev. RNA* **13**, e1724 (2022).
32. Collier, B., Goobar-Larsson, L., Sokolowski, M. & Schwartz, S. Translational inhibition in vitro of human papillomavirus type 16 L2 mRNA mediated through interaction with heterogenous ribonucleoprotein K and poly(rC)-binding proteins 1 and 2. *J. Biol. Chem.* **273**, 22648–22656 (1998).
33. Wan, C. et al.  $\beta$ 2-adrenergic receptor signaling promotes pancreatic ductal adenocarcinoma (PDAC) progression through facilitating PCBP2-dependent c-myc expression. *Cancer Lett.* **373**, 67–76 (2016).
34. Cammack, A. J. et al. A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc. Natl Acad. Sci.* **117**, 10003–10014 (2020).
35. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
36. Parikshak, N. N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
37. Willsey, A. J. et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
38. Harris, J. A. et al. Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. *Front. Neural Circuits* **8**, 76 (2014).
39. Yao, Z. et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222–3241.e26 (2021).
40. Maloney, S. E., Rieger, M. A. & Dougherty, J. D. Identifying essential cell types and circuits in autism spectrum disorders. *Int. Rev. Neurobiol.* **113**, 61–96 (2013).
41. Doller, A. et al. Posttranslational modification of the AU-rich element binding protein HuR by protein kinase Cdelta elicits angiotensin II-induced stabilization and nuclear export of cyclooxygenase 2 mRNA. *Mol. Cell. Biol.* **28**, 2608–2625 (2008).
42. Prechtel, A. T. et al. Expression of CD83 is regulated by HuR via a novel cis-active coding region RNA element. *J. Biol. Chem.* **281**, 10912–10925 (2006).
43. Mulvey, B. & Dougherty, J. D. Transcriptional-regulatory convergence across functional MDD risk variants identified by massively parallel reporter assays. *Transl. Psychiatry* **11**, 1–13 (2021).
44. Gordon, M. G. et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
45. Tewhey, R. et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
46. Rieger, M. A. et al. 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 (2020).
47. Klein, J. C. et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
48. Mulvey, B., Selmanovic, D. & Dougherty, J. D. Sex significantly impacts the function of major depression-linked variants in vivo. *Biol. Psych.* **94**, 466–478 (2023).
49. Beltramo, R. et al. Layer-specific excitatory circuits differentially control recurrent network dynamics in the neocortex. *Nat. Neurosci.* **16**, 227–234 (2013).
50. Vong, L. et al. Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011).
51. Abrantes, A. et al. Gene expression changes following chronic antipsychotic exposure in single cells from mouse striatum. *Mol. Psychiatry* **27**, 2803–2812 (2022).
52. Moudgil, A. et al. Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells. *Cell* **182**, 992–1008 (2020).
53. Doyle, J. P. et al. Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell* **135**, 749–762 (2008).
54. Challis, R. C. et al. Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nat. Protoc.* **14**, 379–414 (2019).
55. Aurnhammer, C. et al. Universal real-time PCR for the detection and quantification of Adeno-associated virus Serotype 2-derived inverted terminal repeat sequences. *Hum. Gene Ther. Methods* **23**, 18–28 (2012).
56. Olivetti, P. R., Lacefield, C. O. & Kellendonk, C. A device for stereotaxic viral delivery into the brains of neonatal mice. *BioTechniques* **69**, 307–312 (2020).
57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
61. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. Younis, I. et al. Rapid-response splicing reporter screens identify differential regulators of constitutive and alternative splicing. *Mol. Cell. Biol.* **30**, 1718–1728 (2010).
63. asd\_3utr\_mpra\_analysis. *figshare* [https://figshare.com/articles/code/asd\\_3utr\\_mpra\\_analysis/24093942/1](https://figshare.com/articles/code/asd_3utr_mpra_analysis/24093942/1) (2023) <https://doi.org/10.6084/m9.figshare.24093942.v1>.

## Acknowledgements

We’d like to thank Bernie Mulvey, Dana King, Rebecca Chase, and the Djuranovic lab for discussions, advice, and edits. 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).

### Author contributions

Conceptualization: T.L., S.P.P., M.A.R., S.J.S., J.D.D.; Methodology: T.L., S.P.P., A.D.F., R.Z.F., M.A.R., D.S., S.S.; Software: T.L., S.P.P., R.Z.F.; Validation: Y.K.S., M.J.K.; Formal Analysis: T.L., S.P.P., A.D.F., R.Z.F., S.B.F., J.Y.A.; Investigation: T.L., S.P.P., A.D.F., R.Z.F., M.A.R., D.S., S.S., Y.K.S., M.J.K., A.F.A.L.; Resources: J.Y.A., S.J.S.; Data Curation: T.L., S.P.P., A.D.F., S.B.F., J.Y.A.; Writing - Original Draft: T.L., R.Z.F., J.D.D.; Writing - Reviewing & Editing: T.L., A.D.F., B.A.C., J.D.D.; Visualization: T.L., S.P.P., A.D.F., R.Z.F., D.S.; Supervision: B.A.C., J.D.D.; Project Administration: B.A.C., J.D.D.; Funding Acquisition: B.A.C., J.D.D.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-05483-w>.

**Correspondence** and requests for materials should be addressed to Joseph D. Dougherty.

**Peer review information** *Communications Biology* thanks Tahsin Stefan Barakat and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: George Inglis.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023