An *in vivo* massively parallel platform for deciphering tissue-specific regulatory function

Authors: Ashley R. Brown[1,3], Grant A. Fox[1,7], Irene M. Kaplow[1,3], Alyssa J. Lawler[2,3,5], BaDoi N. Phan[1,3,4], Morgan E. Wirthlin[1,3,6], Easwaran Ramamurthy[1,3], Gemma E. May[2], Ziheng Chen[2], Qiao Su[1,3], C. Joel McManus[2], and Andreas R. Pfenning[1,2,3]

Affiliations: Carnegie Mellon University Departments of [1]Computational Biology and [2]Biology and [3]Neuroscience Institute; [4]Medical Scientist Training Program, University of Pittsburgh School of Medicine; [5]Current affiliation: Stanley Center for Psychiatric Research, Broad Institute; [6]Current affiliation: Allen Institute for Brain Science; [7]Current affiliation: College of Medicine, University of Kentucky

Abstract

Genetic studies are rapidly identifying non-protein-coding human disease-associated loci. Understanding the regulatory mechanisms underlying these loci remains a challenge because the causal variants and the tissues in which they act are often unclear. Massively parallel reporter assays (MPRAs) have the potential to link differences in genome sequence, including genetic variants, to tissue-specific regulatory function. Although MPRA and similar technologies have been widely adopted in cell culture, there have been several barriers to widespread use in animals. We overcome these challenges with a new whole-animal MPRA (WhAMPRA), where systemic intravenous AAV effectively transduces the plasmid MPRA library to mouse tissues. Our WhAMPRA approach revealed models of tissue-specific regulation that generally match machine learning model predictions. In addition, we measured the regulatory effects of disrupting MEF2C transcription factor binding sites and impacts of late onset Alzheimer's disease-associated genetic variations. Overall, our WhAMPRA technology simultaneously determines the transcriptional functions of hundreds of enhancers *in vivo* across multiple tissues.

Main Text

Transcriptional regulation, a process in which non-coding enhancer sequences play a major role, is a key component of specifying both cell type identity and phenotypic diversity [1–4]. In neural tissue, gene regulatory processes are essential for organizing the range of highly interconnected and regionally specialized cell types that must synchronize their activity to produce behavior [5]. Transcription is largely regulated by enhancers, distal non-coding sequences that are highly tissue-specific relative to proximal promoters [6]. Recent progress in experimental technology has allowed direct profiling of open chromatin, a component of the "epigenomic" (i.e. gene regulatory) landscape, both at the level of tissues as well as individual cell types [7–11]. In parallel, advances in computational methods have enabled the development of sophisticated algorithms, trained on these open chromatin datasets, for predicting the regulatory activity of enhancers from genome sequence alone [12–14].

High-throughput reporter assays [15] can be used to experimentally validate and refine the predictions of machine learning models of gene regulatory activity by measuring the impact of genetic differences across species on regulatory element activity *in vivo*. Chief among these methods is the massively parallel reporter assay (MPRA) [16–20], which involves creating a library of thousands of distinct plasmids, each of which contains a custom-synthesized candidate enhancer element that controls the expression of a unique barcode in conjunction with a minimal promoter (**Fig. 1A**). MPRA libraries can be constructed using enhancer capture as well as other methods [21–23]. In these studies, the expression of these unique plasmid barcodes as RNA, which can be measured in parallel by complementary DNA (cDNA) amplicon sequencing, reflects the transcriptional activity of the corresponding enhancer in the particular cells into which the library has been introduced.

The ability to effectively deliver plasmid libraries of hundreds-to-thousands of candidate enhancers to cultured cells has allowed for detailed, quantitative descriptions of how subtle differences in genome sequence relate to differences in cell type-specific gene regulation. Studies have used reporter assay techniques to measure the effects of single nucleotide polymorphisms (SNPs) identified from expression quantitative trait loci studies (eQTLs) [19,20], SNPs from genome-wide association studies (GWAS) [24–26], and human lineage-specific mutations [27,28]. High-throughput reporter assays have also been adapted to study gene regulation in cultured neurons [18,29]. However, cultured neurons fail to capture the full complexity of highly connected and regionally specialized *in vivo* neural tissue [30,31].

MPRAs have also been adapted to study gene regulation in the brain *in vivo*, but the current assays lack the sensitivity to detect subtle differences in activity from disrupting individual transcription factor binding sites or SNPs. Both electroporation and adeno-associated viruses (AAVs) have been used to deliver MPRA libraries to the mouse brain [23,32–34]. Although these studies have the sensitivity to measure the tissue or even cell type-specificity of enhancers [35], limitations in the number of cells receiving the libraries has made it impossible to detect the impact of genetic variants on brain gene regulation [33,36].

To this end, we developed a whole-animal massively parallel reporter assay (WhAMPRA), which combines a custom, highly modular plasmid with the AAV-PHP.eB virus [37] to deliver the plasmid library containing our reporter assay to a broad range of tissues within a single animal with high reproducibility (**Fig. 1**). With a brief, minimally invasive intravenous injection rather than direct injections, this serotype enters the brain by crossing the blood-brain barrier, increasing the breadth and the consistency of expression across the tissue. This also enables direct comparisons within a single mouse of enhancer activity between different brain regions and the brain versus other tissues. We demonstrate that WhAMPRA can measure the effects of synthetic disruptions of candidate transcription factor binding sites and naturally occurring human variants on tissue-specific enhancer activity.

Results:

WhAMPRA libraries are successfully delivered to tissues across the mouse

We designed the WhAMPRA plasmid (pAAV-MPRAe) and delivery system to maximize transduction and reproducibility. We designed the vector with an Hsp68 minimal promoter, which can stabilize the enhancer-promoter interaction, assisting the initiation of transcription. We also included an mCherry reporter. We cloned the synthesized candidate enhancer sequences and barcodes downstream of both the minimal promoter and mCherry. Cloning sites within the plasmid backbone allow modular changes to both promoter and enhancer sequences (**Supplementary Fig. 1**).

We first tested the transduction and expression of our reporter system by cloning 3 cross-tissue positive control sequences each with one barcode (**Supplementary Table 1**) into the WhAMPRA plasmid after the mCherry coding sequence to make a small MPRA library (MPRAp) followed by transduction into wildtype adult mouse tissue. We confirmed successful transduction and transcription tissues of interest, including neurons in multiple brain regions, by imaging for mCherry (**Fig. 1B**).

We then designed a library of 642 enhancers each paired with 20 unique barcodes (MPRAi) to test the ability of WhAMPRA to relate differences in genome sequence to regulatory differences at multiple levels (**Fig. 1A and Supplementary Table 2**). First, we selected a set of expected positive and negative controls from brain, liver, and immune cells [17,18] based on their measured regulatory activity in previous MPRA experiments. To improve the overall signal from mouse brain, we also added 144 candidate enhancers based on mouse cortex H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) regions and their orthologs across species (see methods). To test the ability of the assay to detect the impact of disrupting transcription factor binding sites, we also synthesized versions of 28 sequences that are known to bind the transcription factor MEF2C, which has been implicated in transcriptional regulation in multiple brain regions [38,39] as well as Alzheimer's disease (AD) predisposition [40]. Finally, we synthesized different versions of a set of 27 candidate enhancers with both the risk and the non-risk alleles of candidate regulatory AD-associated variants from GWAS [41].

We cloned the candidate enhancer sequences into the plasmid backbone and then delivered the libraries into the brains of mice using retro-orbital injection of AAV-PHP.eB [37] as described in Lawler *et al.* [42] (**Fig. 1A**). We collected tissue and sectioned from the liver, the frontal cortex, and the striatum. Imaging nuclear mCherry relative to NeuN (brain) or DAPI (liver) revealed widespread systemic transduction (**Fig. 1B**). Given that not every candidate enhancer is active in every tissue and enhancer activity varies, exact transduction rates are better measured by sequencing of plasmid DNA counts rather than by imaging.

We next sought to catalog the transduction within our library by using DNA sequencing to measure the levels of unique barcodes in cells across tissues. We injected libraries into 8 mice and collected samples from the liver, the primary motor cortex (M1), a larger piece of tissue from the rest of the frontal cortex (referred to as cortex), the hippocampus, and the striatum. We also collected a few samples of tissue from the heart, kidneys, testes, and ovaries. To compare *in vivo* and cell culture technologies, we also transfected our library in the microglia-like HMC3 cell line. The use of HMC3 also allowed us to study the potential function of MEF2 binding sites and

AD-associate genetic variants, which have both been linked to microglia [43,44]. Measurements of the plasmid DNA from each sample were highly correlated (Spearman Rho ranging from 0.737 to 0.991, median 0.951; **Supplementary Fig. 2A and Supplementary Table 3**). Most samples had a high proportion of barcodes detected at the DNA level (**Supplementary Figure 2B**). These results confirm widespread transduction across mouse tissues.

WhAMPRA measures the tissue-specificity of candidate enhancers

The estimated enhancer activity levels suggest that WhAMPRA can reliably measure regulatory activity across different tissues. We used RNA-Seq to measure barcode RNA expression at the RNA levels in HMC3 cells, the liver, the M1, cortex, the hippocampus, and the striatum (**Fig. 1A, right and Supplementary Table 4, 5**). The RNA barcodes showed less reproducibility across samples than the DNA barcodes, likely due to the tissue-specificity of gene regulation (Spearman Rho ranging from 0.366 to 0.995, median 0.629; **Supplementary Fig. 2C and Supplementary Table 3**). Most samples had a high proportion of barcodes detected at the RNA level, but samples with the highest levels came from brain, HMC3, and liver, the tissues for which the enhancers were designed (**Supplementary Figure 2D**).

We used the barcode RNA counts relative to the barcode DNA counts to estimate enhancer activity with MPRAnalyze [45]. We found that the candidate enhancers, which include positive controls, had a strong tendency to be expressed relative to the negative control sequences (**Fig. 2A**). The strongest enrichments for active enhancers came from the brain, where most of the candidate enhancers were expected to be active. Quality control metrics were consistent between the *in vivo* and *in vitro* version of the experiment (**Supplementary Table 3**). For example, RNA:DNA ratios from the HMC3 cells (**Fig. 2B**) showed a similar spread to the RNA:DNA ratios from brain tissue like cortex (**Fig. 2C**).

The activity of the control enhancers provides evidence that WhAMPRA is able to identify tissue-specific patterns. Relative to the negative control candidate enhancers, the set of enhancers active in both HEPG2 (liver-like) and K562 (immune) cells [17] showed a nominal trend toward expression in HMC3 cells (one-sided t-test p=0.078), the liver (one-sided t-test p=0.012), and the brain tissues (one-sided t-test p=0.019; **Fig. 2D**). In contrast, the HEPG2-specific enhancers tended to be transcribed in only the liver (one-sided t-test p=0.039). The cortical control enhancers showed the highest regulatory activity in the brain (one-sided t-test p=0.00025; **Fig. 2D**). Overall, the expression patterns of the positive and negative control candidate enhancers matched expectations. Thus, our experimental design identifies tissue-specific differences in enhancer activity in living mice.

The patterns of expression across all enhancers also provide evidence that the WhAMPRA can capture tissue-specific gene regulation. There was strong correlation between enhancer activity across different brain tissues (Spearman Rho 0.348 to 0.433) but little correlation between brain and liver enhancer activity (Spearman Rho 0.0018 to 0.0971; **Fig. 2E**). The microglia-like cell line, HMC3, showed similarities to both brain and liver tissue (Spearman Rho 0.225 to 0.407) (**Fig. 2E**). To ensure that the enhancer activity we were measuring was related to the tissue-specific regulatory code, we compared the measured activity across all enhancers to machine learning model predictions of open chromatin [12,14], which are correlated with enhancer activity [6] (**Supplementary Table 6**). The predictions of machine learning models trained on brain tissue open chromatin had significant correlations with the WhAMPRA-measured enhancer activity in brain tissue (Spearman Rho 0.121 to 0.183; p from $9.39x10^{-3}$ to $7.90x10^{-5}$) but not liver tissue (Spearman Rho =-0.0117, -0.00370; **Fig. 2F**). Reciprocally, the predictions from a machine learning model trained on liver open chromatin were correlated with enhancer activity in liver (Spearman Rho=0.158; $P=6.68x10^{-4}$) but not brain (Spearman Rho from -0.0112 to 0.0341; **Fig. 2F**). These results show that WhAMPRA identifies tissue-specific signatures of transcriptional enhancers in live mice.

WhAMPRA detects enhancer disruptions from transcription factor binding sites and SNPs

To determine whether WhAMPRA could detect how disruption in transcription factor binding site motifs disrupts enhancer activity *in vivo*, we designed a set of 28 candidate enhancer sequences based on binding the MEF2C transcription factor in mouse cortex (see methods). We also created versions of each enhancer where the transcription factor binding site itself was shuffled, along with an additional version where the motif together with the surrounding 5 nucleotides were shuffled (**Fig. 3A**). The non-disrupted MEF2 motif-containing enhancers showed the strongest activity in brain tissue but also showed some activity in HMC3 cells, consistent with MEF2C's function in both brain and microglia [38,43,46] and its lack of expression in the liver (**Fig. 2D**) [47,48]. Consistent with that observation, we found that disrupting the MEF2C transcription factor binding sites had no significant effect in liver tissue (p>0.1; **Fig. 3B, top**) but significantly decreased enhancer activity in HMC3 cells and in cortical tissue (one-tailed t-test p-value = 0.002; **Fig. 3B, middle and bottom**). Notably, the extent to which enhancer activity was disrupted by shuffling the MEF2C binding site was highly correlated with the original baseline expression of the enhancer (**Fig. 3C; Supplementary Table 7**; Rho=0.88; $p=9.3x10^{-7}$). This suggests that the cases where disrupting the MEF2C transcription factor binding site has no effect are cases where the enhancer itself is not active in the assay rather than cases where the MEF2C binding site is not important for enhancer activity.

Next, we used WhAMPRA to measure the impact that 27 Alzheimer's Disease GWAS-derived SNPs have on regulatory activity. We synthesized both the risk and non-risk allele of non-coding SNPs implicated in an AD genome-wide association study [41]. For the subset of those enhancers where the SNP disrupted an important transcription factor binding site [49], we also included a version of the enhancer where the entire transcription factor binding site was shuffled. There were eight enhancers where the two alleles showed differential activity across the HMC3 and brain samples, two of which (rs6498140, rs10991386) were confirmed by the motif disruption (**Fig. 3D**; **Supplementary Table 8**).

The strongest evidence for the impact of a candidate AD-associated SNP on enhancer activity is for rs6498140, which is proximal to the gene *CLEC16A*. The alternate allele has the highest regulatory activity in both brain tissue and in HMC3 (**Fig. 3E**). The alternate allele creates a MEF2 transcription factor binding site motif in the enhancer (**Fig. 3F**), which is consistent with

MEF2 being an active transcription factor in both brain and microglia [43]. Notably, the SNP rs6498140 displays marginal GTEx eQTL associations in several tissues including the frontal cortex, where the alternate allele correlates with higher expression of *CLEC16A* [50]Thus, our WhAMPRA both validates and expands the understanding of genetic variation and gene regulation in AD *in vivo* at relevant tissues.

Discussion

While MPRA technology has been instrumental in linking genome sequence to regulatory function, thus far it has been used primarily in cell culture. Here, we developed WhAMPRA, a massively parallel reporter assay with the ability to measure regulatory activity across tissues *in vivo*. In concert with other experimental design choices, the delivery of the MPRA library into cells using systemic AAV transduction helps us achieve the necessary sensitivity to detect the tissue-specific regulatory effects of mutating transcription factor binding sites and individual nucleotides. Systemic delivery reduces local inflammation caused by intraparenchymal injection and thereby enables the study of disease processes with inflammatory components [51,52]. Furthermore, systemic delivery increases the throughput of delivering MPRA to multiple tissues of interest in the same experimental animal with one brief, minimally invasive procedure.

A comparison of regulatory activity in brain and liver demonstrates that WhAMPRA can detect highly tissue-specific regulatory activity. We then use candidate enhancers known to bind the MEF2C transcription factor, which is active in the brain in microglia, to show that the assay can detect the impact of disrupting MEF2C transcription factor binding sites on regulatory activity. We also use a set of candidate AD-associated mutations to show that the assay is sensitive enough to detect the impact of disrupting individual SNPs on enhancer activity.

Although we are able to detect tissue- and allele-specific effects, there are still limitations in WhAMPRA's *in vivo* technology. In contrast to the hundreds of enhancers we profile, current cell culture reporter assays can provide quantitative, cell line-specific information across thousands of enhancers [17,27,53]. Efficient transduction of tissue is likely the current limitation in the technology. Furthermore, any AAV tropism present in PHP.eb or another systemic delivery system will also be reflected in the measured enhancer activity. For example, gene regulatory programs active in brain microglia are not likely to be captured in our WhAMPRA experiment due to the bias that PHP.eb has for neurons and other glial cells [37].

As AAV technology improves, we expect WhAMPRA to become more flexible. We designed our library to have active enhancers in brain, liver, and immune cells, but we detect transduction in heart, kidney, and other other tissues. New AAV variants are being designed that would allow WhAMPRA libraries to be targeted to specific cell subtypes [54,55] and even non-human primate tissue [56]. As transduction efficiency improves, WhAMPRA could be paired with a methodology to isolate individual cell types [9,42] or even single-cell profiling [32].

GWAS and whole-genome sequencing studies are identifying an increasing number of candidate regulatory variants underlying the predisposition to complex traits. Fine-mapping and functional characterization of those variants is an important step in connecting genetic predisposition to disease pathophysiology. While *in vitro* high-throughput reporter assays have provided an avenue for high-throughput functional characterization in human cells, there may be genetic variants that act in a cell type or cellular environment not able to be captured *in vitro*. Complementary to *in vitro* versions of MPRA, WhAMPRA enables high throughput functional characterization across tissues of a behaving organism. This enables tissue-specific regulatory effects to be measured in animal models of disease, including non-traditional model organisms.

Acknowledgements:

One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. We thank Drs. Viviana Gradinaru and Benjamin Deverman for early use of their systemic brain-transducing viral capsid. We thank Dr. Jesse Gray and the members of his lab for their discussion of early versions of WhAMPRA. We thank Dr. Oliver Schlüter and Corinne Schneider from the University of Pittsburgh for their training and discussions about AAV production. We thank Tae Yoon Park and other current and former members of the Pfenning lab as well as the CMU Neuroscience community, especially Drs. Alison Barth, Aryn Gittis, and Sandra Khulman, for their discussions and feedback throughout this project. MEW was funded by the Carnegie Mellon BrainHub Postdoctoral Fellowship. IMK was funded by the Carnegie Mellon Computational Biology Department Lane Fellowship. BNP was funded by the NIH National Institutes on Drug Abuse (F30DA053020). AJL was funded by the NSF (DGE1745016). ARP was funded by the NIH National Institutes on Drug Abuse (DP1DA046585) and the CURE Alzheimer's Fund.

Figures:

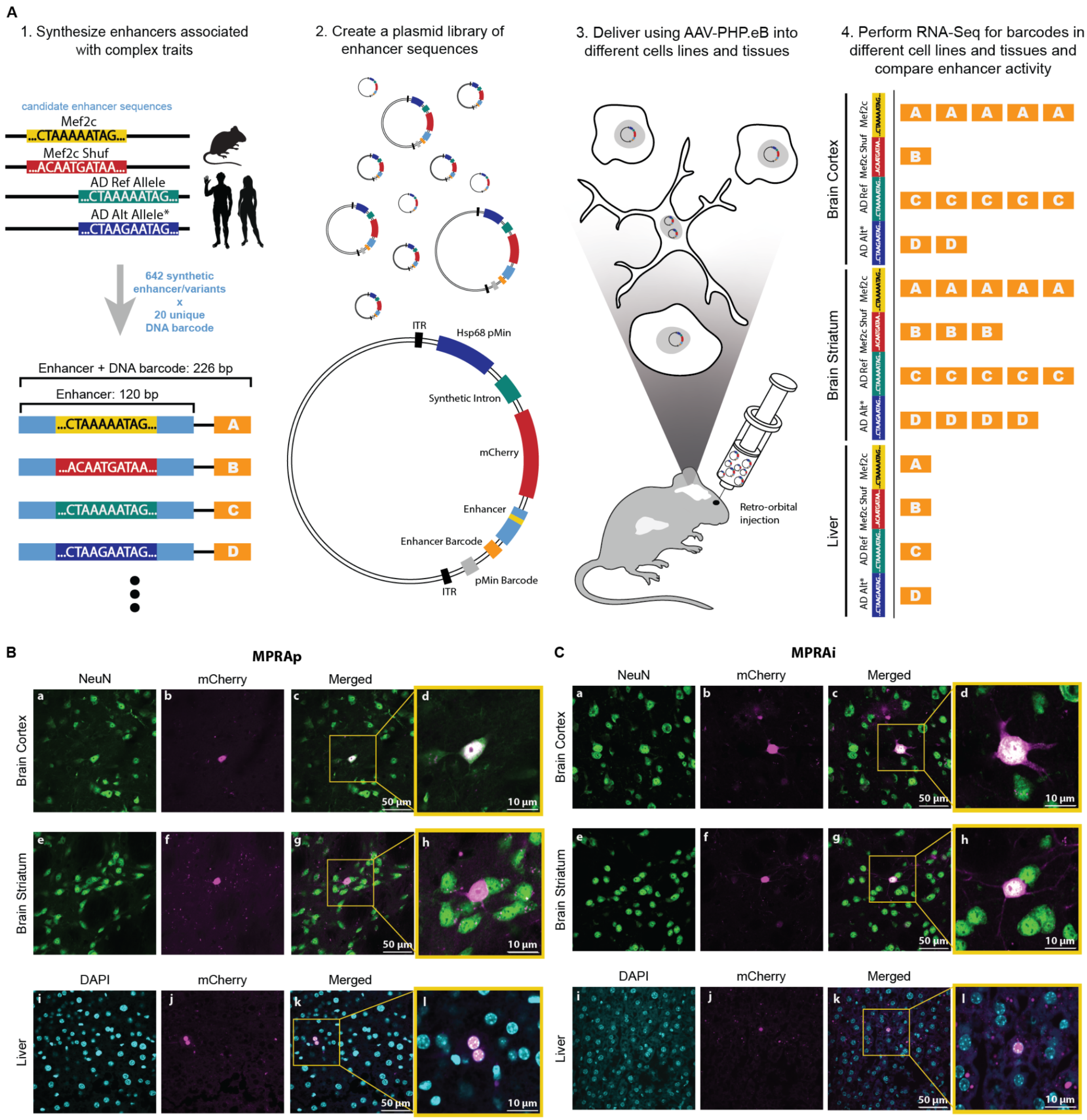

Figure 1 - WhAMPRA tests effects of transcription factor binding and single nucleotide variations on transcriptional regulation. (A) 1. The library is designed to study complex traits consisting of 642 enhancers and variants, each with 20 unique barcodes (i.e. MEF2C motifs and shuffled versions of these motifs, as well as reference and alternative alleles for AD-associated SNPs). 2. The oligos are synthesized and cloned into plasmids. 3. The plasmid library is packaged into the PHP.eB AAV serotype and delivered into a mouse via retro-orbital injection. 4. The activity of candidate enhancers in multiple brain regions and tissues is measured using

RNA levels of the bar-codes. (B) mCherry expression from MPRAp (cross-tissue positive controls). Shown is mCherry (magenta) compared to NeuN expression (green) in the brain cortex (panels a-d) and brain striatum (panels e-h) from a C57Bl/6J mouse. mCherry (magenta) is also compared to DAPI (blue) expression in the liver (panels i-l) from a C57Bl/6J mouse. (C) mCherry expression from MPRAi (MPRA library of 642 enhancers/variants). Shown is mCherry (magenta) compared to NeuN expression (green) in the brain cortex (panels a-d) and brain striatum (panels e-h) from C57Bl/6J mouse. mCherry (magenta) is also compared to DAPI (blue) expression in the liver (panel i-l) from a C57Bl/6J mouse.

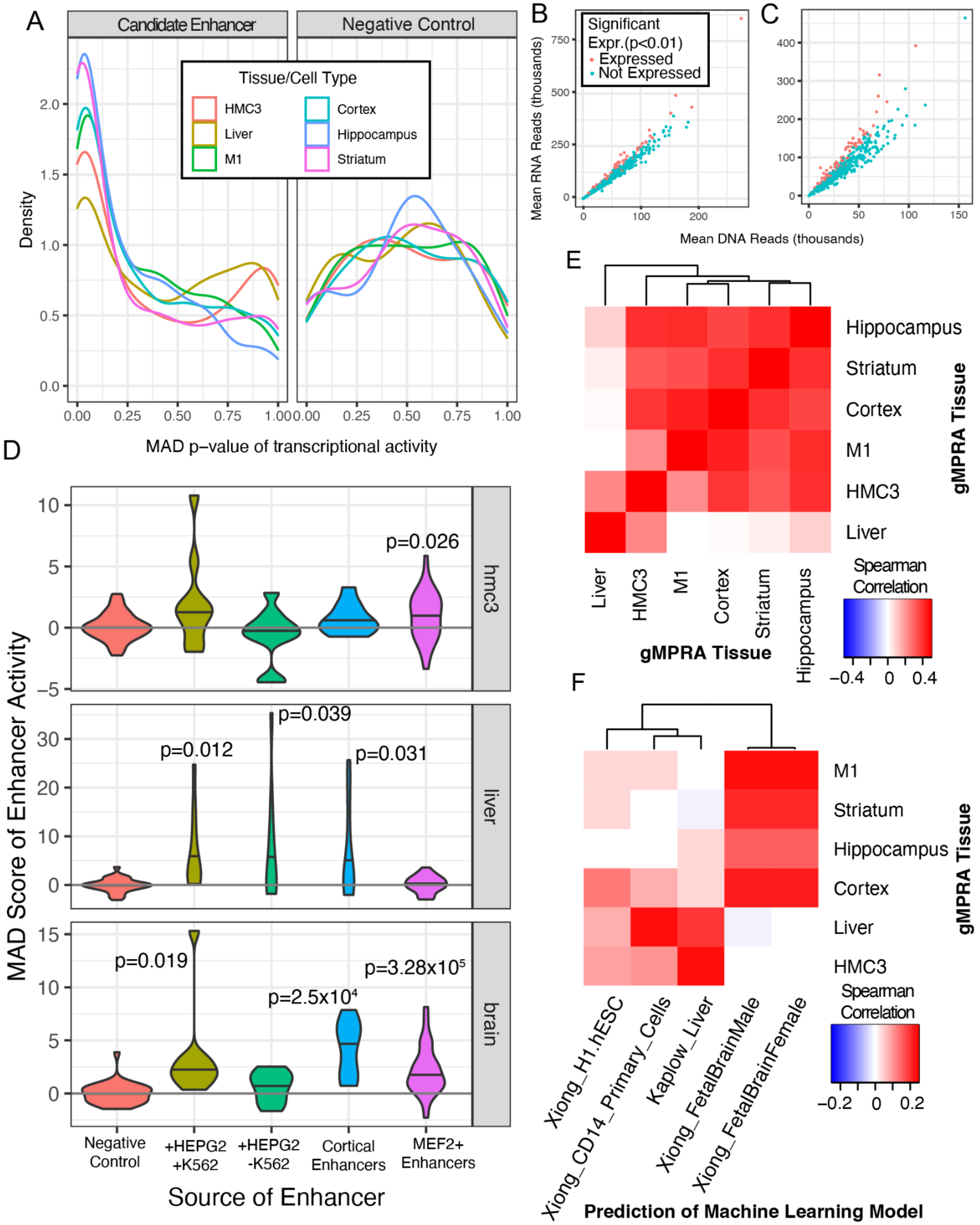

Figure 2 - WhAMPRA captures tissue-specific signatures of gene regulation *in vivo*. (A) The frequency of p-values is displayed using a density plot across the candidate enhancers and positive controls (left) relative to the negative controls (right). The ratio of DNA reads to RNA reads, which roughly corresponds to transcriptional activity, is plotted for HMC3 cultured cells

(B) and for cortical tissue (C). The mean across all samples for that tissue is used. (D) The MAD score is displayed as a violin plot for the likely positive and negative control enhancers gleaned from other MPRA experiments. (E) Spearman's Rho is calculated across the estimated transcription rate, alpha, of all enhancers for each pairwise tissue comparison. (F) Spearman's Rho is calculated between the estimated transcription rate, alpha, of all enhancers and the prediction of open chromatin levels calculated by the convolutional neural network models.

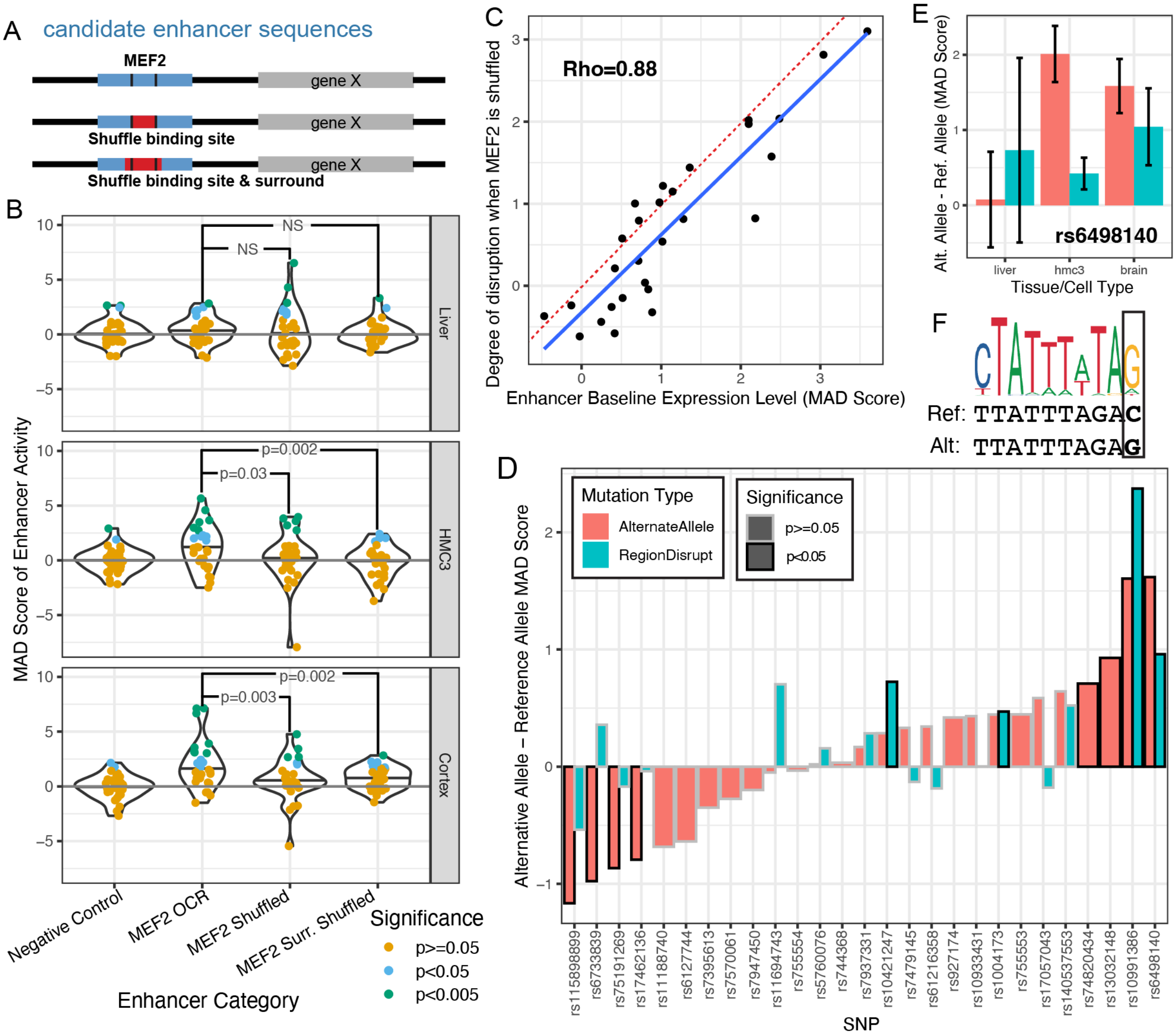

Figure 3 - WhAMPRA detects enhancer differences due to MEF2C binding site disruption and candidate Alzheimer's disease SNPs. (A) The experimental design of how MEF2C was systematically disrupted at candidate enhancers with binding sites. (B) The MAD score of enhancer activity is compared between negative control enhancers and different versions of candidate MEF2C enhancers. Each enhancer is colored based on its nominal significance of transcription relative to the population of negative controls (MAD p-value). (C) The MAD score of baseline enhancer expression is compared to the difference between the baseline enhancer expression and the average expression across the two instances of MEF2C shuffling. (D) The MAD score of the enhancer activity from the reference allele is compared to the alternate allele (red) and the sequence with a shuffled local transcription factor binding site (blue). (E) The MAD score of the alternative allele for SNP rs6498140 is compared to the MAD score of the reference allele for the candidate AD-associated enhancer. (F) The motif logo for a discovered MEF2 transcription factor binding site is visualized above the reference and alternative allele for rs6498140.

References:

1. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
2. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007).
3. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).
4. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).
5. Goodman, J. V. & Bonni, A. Regulation of neuronal connectivity in the mammalian brain by chromatin remodeling. *Curr. Opin. Neurobiol.* **59**, 59–68 (2019).
6. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
7. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
8. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
9. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86**, 1369–1384 (2015).
10. Lawler, A. J. *et al.* Cell Type-Specific Oxidative Stress Genomic Signatures in the Globus Pallidus of Dopamine-Depleted Mice. *J. Neurosci.* **40**, 9772–9783 (2020).
11. Bryois, J. *et al.* Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat. Commun.* **9**, 1–15 (2018).

12. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
13. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
14. Kaplow, I. M. *et al.* Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. *BMC Genomics* **23**, 291 (2022).
15. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
16. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
17. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
18. Nguyen, T. A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
19. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
20. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).
21. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 1–15 (2018).
22. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
23. Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* **26**, 238–255 (2016).
24. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).

25. Chaudhri, V. K., Dienger-Stambaugh, K., Wu, Z., Shrestha, M. & Singh, H. Charting the cis-regulome of activated B cells by coupling structural and functional genomics. *Nat. Immunol.* **21**, 210–220 (2020).

26. Myint, L. *et al.* A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **183**, 61–73 (2020).

27. Jagoda, E. *et al.* Detection of Neanderthal Adaptively Introgressed Genetic Variants That Modulate Reporter Gene Expression in Human Immune Cells. *Mol. Biol. Evol.* **39**, (2022).

28. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proceedings of the National Academy of Sciences* vol. 118 Preprint at https://doi.org/10.1073/pnas.2007049118 (2021).

29. Girskis, K. M. *et al.* Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**, 3239–3251.e7 (2021).

30. Gordon, J., Amini, S. & White, M. K. General overview of neuronal cell culture. *Methods Mol. Biol.* **1078**, 1–8 (2013).

31. Lawler, A. J. *et al.* Machine learning sequence prioritization for cell type-specific enhancer design. *Elife* **11**, e69571 (2022).

32. Hrvatin, S. *et al.* A scalable platform for the development of cell-type-specific viral drivers. *Elife* **8**, (2019).

33. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *Elife* **10**, (2021).

34. Warren, T. L., Lambert, J. T. & Nord, A. S. AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *J. Vis. Exp.* (2022) doi:10.3791/62650.

35. Blankvoort, S., Witter, M. P., Noonan, J., Cotney, J. & Kentros, C. Marked Diversity of Unique Cortical Enhancers Enables Neuron-Specific Tools by Enhancer-Driven Gene Expression. *Curr. Biol.* **28**, 2103–2114.e5 (2018).

36. Mulvey, B., Lagunas, T., Jr & Dougherty, J. D. Massively Parallel Reporter Assays: Defining

Functional Psychiatric Genetic Variants Across Biological Contexts. *Biol. Psychiatry* **89**, 76–89 (2021).

37. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).
38. Harrington, A. J. *et al.* MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *Elife* **5**, e20059 (2016).
39. Chen, Y. C. *et al.* Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. *Nat. Neurosci.* **19**, 1513–1522 (2016).
40. Karch, C. M. & Goate, A. M. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol. Psychiatry* **77**, 43–51 (2015).
41. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
42. Lawler, A. J. *et al.* Machine learning sequence prioritization for cell type-specific enhancer design. *eLife* vol. 11 Preprint at https://doi.org/10.7554/elife.69571 (2022).
43. Deczkowska, A. *et al.* Mef2C restrains microglial inflammatory response and is lost in brain ageing in an IFN-I-dependent manner. *Nat. Commun.* **8**, 717 (2017).
44. Gjoneska, E., Pfenning, A. R., Mathys, H. & Quon, G. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* (2015).
45. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
46. Telese, F. *et al.* LRP8-Reelin-Regulated Neuronal Enhancer Signature Underlying Learning and Memory Formation. *Neuron* **86**, 696–710 (2015).
47. The Human Protein Atlas.
48. Baldarelli, R. M. *et al.* The mouse Gene Expression Database (GXD): 2021 update. *Nucleic Acids Res.* **49**, D924–D931 (2021).
49. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell

types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–81 (2016).

50. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
51. Seney, M. L. *et al.* Transcriptional Alterations in Dorsolateral Prefrontal Cortex and Nucleus Accumbens Implicate Neuroinflammation and Synaptic Remodeling in Opioid Use Disorder. *Biol. Psychiatry* **90**, 550–562 (2021).
52. Wyss-Coray, T. & Rogers, J. Inflammation in Alzheimer Disease—A Brief Review of the Basic Science and Clinical Literature. *Cold Spring Harb. Perspect. Med.* **2**, (2012).
53. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).
54. Bryant, D. H. *et al.* Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
55. Öztürk, B. E. *et al.* scAAVengr, a transcriptome-based pipeline for quantitative ranking of engineered AAVs with single-cell resolution. *Elife* **10**, (2021).
56. Goertsen, D. *et al.* AAV capsid variants with brain-wide transgene expression and decreased liver targeting after intravenous delivery in mouse and marmoset. *Nat. Neurosci.* **25**, 106–115 (2022).
57. Benoist, C. & Chambon, P. In vivo sequence requirements of the SV40 early promotor region. *Nature* **290**, 304–310 (1981).
58. Thomsen, D. R., Stenberg, R. M., Goins, W. F. & Stinski, M. F. Promoter-regulatory region of the major immediate early gene of human cytomegalovirus. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 659–663 (1984).
59. Schlabach, M. R., Hu, J. K., Li, M. & Elledge, S. J. Synthetic design of strong promoters. *Proceedings of the National Academy of Sciences* vol. 107 2538–2543 Preprint at https://doi.org/10.1073/pnas.0914803107 (2010).

60. Yaguchi, M. *et al.* Characterization of the Properties of Seven Promoters in the Motor Cortex of Rats and Monkeys After Lentiviral Vector-Mediated Gene Transfer. *Human Gene Therapy Methods* vol. 24 333–344 Preprint at https://doi.org/10.1089/hgtb.2012.238 (2013).

61. Carullo, N. V. N. & Day, J. J. Genomic Enhancers in Brain Health and Disease. *Genes* vol. 10 43 Preprint at https://doi.org/10.3390/genes10010043 (2019).

62. Pai, E. L.-L. *et al.* Maf and Mafb control mouse pallial interneuron fate and maturation through neuropsychiatric disease gene regulation. *Elife* **9**, e54903 (2020).

63. Mitchell, A. C. *et al.* MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. *Mol. Psychiatry* **23**, 123–132 (2018).

64. Lee, J. W., Foo, C. S., Kim, D., Boley, N. & Kundaje, A. ATAC-Seq / DNase-Seq Pipeline. https://github.com/kundajelab/atac_dnase_pipelines (2017).

65. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

66. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

67. Srinivasan, C. *et al.* Addiction-associated genetic variants implicate brain cell type- and region-specific cis-regulatory elements in addiction neurobiology. (2020) doi:10.1101/2020.09.29.318329.

68. Stamatoyannopoulos, J. A. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**, 418 (2012).

69. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

70. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic

features. *Bioinformatics* **26**, 841–842 (2010).

72. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

73. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

74. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

75. Lee, J. W., Boley, N. & Kundaje, A. *AQUAS TF and histone ChIP-seq pipeline*. (2016).

76. Marinov, G. K., Kundaje, A., Park, P. J. & Wold, B. J. Large-Scale Quality Analysis of Published ChIP-seq Data. *G3: Genes|Genomes|Genetics* **4**, 209–223 (2014).

77. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

78. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

79. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).

80. Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat. Neurosci.* **19**, 494–503 (2016).

81. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

82. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

83. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

84. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).

85. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11484–11489 (2003).

86. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

87. Partha, R. *et al.* Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**, e25884 (2017).

88. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

89. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

90. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757–762 (2010).

91. Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).

92. Pavlovich, S. S. *et al.* The Egyptian Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. *Cell* **173**, 1098–1110.e18 (2018).

93. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

94. Wirthlin, M. *et al.* A Modular Approach to Vocal Learning: Disentangling the Diversity of a Complex Behavioral Trait. *Neuron* **104**, 87–99 (2019).

95. Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).

96. Hannula-Jouppi, K. *et al.* The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genet.* **1**, e50 (2005).

97. Newbury, D. F. *et al.* CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am. J. Hum. Genet.* **85**, 264–272 (2009).

98. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213–217 (2009).

99. König, I. R. *et al.* Mapping for dyslexia and related cognitive trait loci provides strong evidence for further risk genes on chromosome 6p21. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **156B**, 36–43 (2011).

100. Horn, D. *et al.* Identification of FOXP1 deletions in three unrelated patients with mental retardation and significant speech and language deficits. *Hum. Mutat.* **31**, E1851–60 (2010).

101. Filges, I. *et al.* Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel--Giedion syndrome. *J. Med. Genet.* **48**, 117–122 (2011).

102. Roeske, D. *et al.* First genome-wide association scan on neurophysiological endophenotypes points to trans-regulation effects on SLC2A3 in dyslexic children. *Mol. Psychiatry* **16**, 97–107 (2011).

103. Bacon, C. & Rappold, G. A. The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. *Hum. Genet.* **131**, 1687–1698 (2012).

104. Thevenon, J. *et al.* 12p13.33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. *Eur. J. Hum. Genet.* **21**, 82–88 (2013).

105. Amarillo, I. E., Li, W. L., Li, X., Vilain, E. & Kantarci, S. De novo single exon deletion of AUTS2 in a patient with speech and language disorder: a review of disrupted AUTS2 and further evidence for its role in neurodevelopmental disorders. *Am. J. Med. Genet. A* **164A**, 958–965 (2014).

106. Peter, B., Matsushita, M., Oda, K. & Raskind, W. De novo microdeletion of BCL11A is associated with severe speech sound disorder. *Am. J. Med. Genet. A* **164A**, 2091–2096 (2014).

107. Namjou, B. *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric

cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Front. Genet.* **5**, 401 (2014).

108. Turner, S. J. *et al.* GRIN2A: an aptly named gene for speech dysfunction. *Neurology* **84**, 586–593 (2015).

109. Chen, X. S. *et al.* Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. *Sci. Rep.* **7**, 46105 (2017).

110. Senkevich, K. *et al.* Associations of genetic variants in COMT, BDNF, SNCA, MAPT genes with cognitive impairment in Parkinson's disease. (P6.090). *Neurology* **90**, (2018).

111. Eising, E. *et al.* A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol. Psychiatry* **24**, 1065–1078 (2019).

112. Kielb, S., Kisanuki, Y. Y. & Dawson, E. Neuropsychological profile associated with an alpha-synuclein gene (SNCA) duplication. *Clin. Neuropsychol.* 1–12 (2021).

113. Jarvis, E. D. Evolution of vocal learning and spoken language. *Science* **366**, 50–54 (2019).

114. Prat, Y., Taub, M. & Yovel, Y. Vocal learning in a social mammal: Demonstrated by isolation and playback experiments in bats. *Science Advances* vol. 1 Preprint at https://doi.org/10.1126/sciadv.1500019 (2015).

115. Genzel, D., Desai, J., Paras, E. & Yartsev, M. M. Long-term and persistent vocal plasticity in adult bats. *Nat. Commun.* **10**, 3372 (2019).

116. Neuroscience of birdsong. **550**, (2008).

117. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635 (2008).

118. Ramamurthy, E. *et al.* Cell type-specific histone acetylation profiling of Alzheimer's Disease subjects and integration with genetics. *bioRxiv* 2020.03.26.010330 (2020) doi:10.1101/2020.03.26.010330.

119. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

120. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types.

Nature **473**, 43–49 (2011).

121. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2011).

122. Nott, A. *et al.* Brain cell type-specific enhancer-promoter interactome maps and disease - risk association. *Science* **366**, (2019).

123. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, (2015).

124. Deverman, B. E. *et al.* Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34**, 204–209 (2016).

125. Lock, M. *et al.* Rapid, simple, and versatile manufacturing of recombinant adeno-associated viral vectors at scale. *Hum. Gene Ther.* **21**, 1259–1271 (2010).

126. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).

127. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

128. Chollet, F. Keras. https://keras.io (2015).