bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: The Oft-Overlooked Massively Parallel Reporter Assay: Where, When, and Which Psychiatric Genetic Variants are Functional?

Authors:

Bernard Mulvey[1,2,3], Tomas Lagunas[1,2,3], Joseph D. Dougherty[2,3]

[1]Division of Biology and Biomedical Sciences, Washington University School of Medicine, St. Louis, MO, USA
[2]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[3]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

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

CONFLICT OF INTEREST

No authors declare a conflict of interest.

ACKNOWLEDGEMENTS

This work was funded by NIMH grants 1F30MH1116654 and 1R01MH116999, Simons Foundation 571009 and NHGRI grant 1R01HG008687. The *Camk2a* promoter was derived from a plasmid generously gifted by Karl Deisseroth to Washington University's Viral Vector Core**.** We would like to thank Sergej Djuranovic, Ph.D., Barak Cohen, Ph.D., and Cohen lab alumni Dana King, Ph.D., and Brett Maricque, Ph.D. for their collaboration and guidance designing, adapting, and analyzing MPRAs applied to neuropsychiatry. We would also like to thank Idoya Lahortiga, Ph.D. and Luk Cox, Ph.D. curators of Somersault1824 (https://www.Somersault1824.com), for their open-access, Creative Commons BY-NC-SA 4.0 licensed libraries of high-quality biomedicine graphics (especially those from Graphite Life Explorer, ePMV, and Eyewire), adapted for figures in this review. Finally, we would like to thank Mike Vasek for his assistance in editing this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Neuropsychiatric phenotypes have been long known to be influenced by heritable risk factors. The past decade of genetic studies have confirmed this directly, revealing specific common and rare genetic variants enriched in disease cohorts. However, the early hope for these studies—that only a small set of genes would be responsible for a given disorder—proved false. The picture that has emerged is far more complex: a given disorder may be influenced by myriad coding and noncoding variants of small effect size, and/or by rare but severe variants of large effect size, many *de novo*. Noncoding genomic sequences harbor a large portion of these variants, the molecular functions of which cannot usually be inferred from sequence alone. This creates a substantial barrier to understanding the higher-order molecular and biological systems underlying disease risk. Fortunately, a proliferation of genetic technologies—namely, scalable oligonucleotide synthesis, high-throughput RNA sequencing, CRISPR, and CRISPR derivatives—have opened novel avenues to experimentally identify biologically significant variants *en masse*. These advances have yielded an especially versatile technique adaptable to large-scale functional assays of variation in both untranscribed and untranslated regulatory features: Massively Parallel Reporter Assays (MPRAs). MPRAs are powerful molecular genetic tools that can be used to screen tens of thousands of predefined sequences for functional effects in a single experiment. This approach has several ideal features for psychiatric genetics, but remains underutilized in the field to date. To emphasize the opportunities MPRA holds for dissecting psychiatric polygenicity, we review here its applications in the literature, discuss its ability to test several biological variables implicated in psychiatric disorders, illustrate this flexibility with a proof-of-principle, *in vivo* cell-type specific implementation of the assay, and envision future outcomes of applying MPRA to both computational and experimental neurogenetics.

Introduction

Psychiatric diseases are genetically influenced by common and rare heritable variation, as well as by non-inherited, *de novo* mutations. Liability to disease through common variants (allele frequency ≥ 1%) ranges from 10-25% for psychiatric disorders including major depressive disorder (MDD) (1–3) and schizophrenia (1,4). Rare variants have likewise been shown to confer disease risk, particularly in neurodevelopmental and psychotic (5,6) domains. While loci containing both types of variants have been associated with neuropsychiatric phenotypes, two major hurdles have prevented use of this knowledge to better understand disease mechanisms: 1) a large volume of GWAS associations and rare variant discoveries to test for functionality/causality; and 2) the lack of means to predict a variant's functional consequences. These hurdles confound the identification of shared features across functional variants that converge on common phenotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Defining specific causal variant(s): problems of linkage and numbers

The identification of candidate genetic variants with disease associations relies on correlational methodologies. Genome wide association studies (GWAS) identify loci with overrepresented blocks of linked variants, comprising numerous single nucleotide polymorphisms (SNPs). Similarly, family studies (*e.g.*, trio studies) identify many proband-specific (*de novo*) or -enriched (rare, inherited) variants in each patient, though only 1-2 may be causal. Further, while consequences of coding variants (*e.g.,* nonsense mutations) can be reliably predicted, the vast majority of GWAS loci fall outside of transcribed sequences(7). Rare and *de novo* variation, on the other hand, is more broadly distributed across the genome, including un*translated* and translation-regulating sequences, well-illustrated by autism spectrum disorders (ASDs): over 255,106 distinct *de novo* variants (including 142 stop codon gains, 3,402 in untranslated regions (UTRs), and 6,787 upstream gene/promoter variants) were recently identified in 1,902 ASD subjects (8). However, these discovery-oriented approaches are incapable of identifying the minority of associations corresponding to variants with biological/disease-pertinent functions.

Challenges in predicting a variant's functional consequences

Predicting whether and how a noncoding variant is functional is a nontrivial enterprise. The majority of these variants, and their linked neighbors, bear indirect indication(s) of transcriptional regulatory function in annotations such as expression quantitative trait locus (eQTL) associations, chromatin accessibility, or histone marks (7,9,10). However, even within a cell or tissue type, such data are often mutually discordant: one study examining six epigenomic datasets in K562 cells found that 49% of functional regulators did not overlap *any* of the six annotation sets, and another 40% only overlapped one of the six (11). Similarly, the Roadmap Epigenomics Consortium (12) inferred cell type-specific regulatory elements using chromatin marks, but only a minority of GWAS SNPs overlap these elements (except in blood) (13). Therefore, epigenomic data alone are inadequate for predicting functional regulatory variants in a given cell type.

While there is a clear excess of *de novo* variation in coding sequences in ASD and other neurodevelopment disorders, substantial additional burden is estimated to be borne by other elements, such as transcriptional (8) and translational regulators (*e.g.*, splicing 5'/3' untranslated regions; UTRs) (14,15). These regions function to regulate transcript stability and miRNA interactions (16); they may also mediate nuclear retention/export of transcripts in the brain, which has only recently been explored (17)**.** Coding variation carries a minority of heritable risk for schizophrenia,

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bipolar disorder, and attention deficit hyperactivity disorder (18). The occurrence of most disease-linked variation in the least-well understood features of the genome/transcriptome thus obstructs understanding of disease biology.

Collectively, these two problems necessitate **high-throughput assays with functional readout** for putative regulatory regions and variants. Such assays enable identification of functional variants and of the biological and environmental contexts in which they act. This knowledge, in turn, can begin shaping hypotheses regarding the shared mechanisms by which seemingly disparate genetic factors converge on shared phenotypic endpoints.

Here, we will primarily discuss MPRAs in terms of their potential for high-throughput parcellation of genetic studies. This technology relies on pairing genomic features (*e.g.*, an allelic pair of a variant-containing region) to a reporter gene bearing unique, transcribed barcodes, allowing an RNA-level readout of features' activities in a multiplex setting (19,20). This approach can be flexibly implemented allowing study of transcriptional regulators (*e.g.*, promoters), as well as splicing, protein translation, and post-transcriptional regulators. Critically, the potential for MPRAs to dissect roles of variants from neuropsychiatric genetic studies is far from fully realized. Here, we aim to illustrate that MPRAs are a methodologic "low-hanging fruit" for neuropsychiatric genetics. In the first part of this review, we discuss the importance of cellular, biological, and environmental contexts in the design and execution of MPRA, exemplify uses of the approach to date, and discuss complementary/follow-up methods to further validate functional variants nominated by MPRA. In the second part, we discuss multiple features of MPRA that are well-suited to parsing polygenic architecture—both for dissecting linked blocks of multiple functional variants, and for identifying convergent variant mechanisms across the genome.

Part 1: MPRAs for Identification of Sequence Variants with Functional Consequences

MPRAs **quantify activity of putative regulatory elements by coupling them to a reporter gene and counting transcribed, element-specific tags ("barcodes"), using sequencing to determine the ratio of (expressed RNA barcode)/(delivered DNA barcode).** This experimental framework has been applied to human splicing (21–23), RNA editing (24), protein translation (25), UTRs (26–30), and, most broadly, transcriptional (*i.e.*, *cis*-) regulators. MPRAs thus offer a flexible framework to study regulatory phenomenon, including transcription factor (TF) and RNA binding protein (RBP) actions, transcript stability, and ribosome occupancy. MPRAs have been most broadly applied to explore and computationally model "regulatory grammar" of transcriptional regulators: how sequence

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

features such as binding motifs, their abundance, and arrangement affect regulatory capacity (31–38). More recently, these approaches have been applied to identify the transcriptional consequences of SNPs and rare variants (39–45).

As shown in **Figure 1A-B,** a canonical 'enhancer' MPRA utilizes a promoter with candidate elements either immediately upstream or in a 3'UTR (STARRseq) (same approach as **Figure 1D**) (46). Each sequence element is paired to multiple, unique barcodes, which are transcribed into UTRs and sequenced as quantitative readout. Expression—representing transcription or RNA stability—is typically measured as the counts of RNA barcode per encoding DNA barcode (**Figure 1G**). To define active or differentially active elements, expression levels can be normalized to *e.g.*, a minimal-promoter only set of barcodes (31,34,37,47–49), compared between alleles (39–45), or compared to shuffled parent sequence(s) (32,37)

MPRAs also enable study of post-transcriptional regulatory elements. As shown in **Figure 1C and 1D**, the same architecture and RNA/DNA expression metric can be used to assess UTR effects on transcript stability. UTR MPRAs have yet to be implemented to study regulatory variants directly, but have been used to study the regulatory grammar of the ASD/ID-implicated CELF proteins and related RBPs (26,50,51), features conferring transcript stability via 3'UTRs (27,28,30), and 5'UTR influences on translation (52) and prediction of functional 5'UTR variants (29). Across enhancer and UTR MPRAs, several key forms of disease-associated noncoding variation can be assessed for functional consequences using a variety of model systems and delivery approaches (**Figure 1E-F**).

MPRAs Enable Identification of Functional Regulators and Variants In Specific Cellular Contexts

Perhaps the most exciting—if underappreciated—property of MPRAs is the ability to assay regulators of interest in the relevant cell types and conditions. Functional elements both define—and are defined by—a given cell type's unique milieu of expressed TFs, chromatin modifiers, miRNAs and RBPs, which in turn specify regulatory element functionality. The importance of cell type is strikingly illustrated by a *cis*-regulatory MPRA of random sequences performed in parallel using U87 glioblastoma and neural progenitor cells. Comparison of the most active enhancers across the two cell types yielded entirely different motifs and sequence features correlated with expression (49). This **sensitivity to cellular context** highlights that **MPRAs can be designed to assay and compare genetic regulation in cell types known—or predicted to—mediate disease.** This also highlights the fact that careful consideration needs to be given to the appropriate cellular context when

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

designing assays for psychiatric genetics: the array of variant-interacting TFs and RBPs expressed in neural cells may not be present in convenient cell lines (*e.g.,* HeLa or HEK-293).

While cell type overwhelmingly influences outcomes in regulatory assays, additional conditions could equally alter outcomes (**Figure 2**). Age, sex, pharmacology, and environment (*e.g.*, stress)—all, like cell type, can shape or reshape regulatory activity. This capacity of MPRAs has recently been applied in neurogenetics—specifically in determining the temporal patterns of regulatory element activity over the course of differentiation of human neural progenitor cell (NPC) differentiation to neurons (53).

In addition to context-specific element activity, MPRAs can also identify context-specific functional *variants*. This follows from findings that the influence of genomic variants on epigenomic marks is also cell type dependent, even within a developmental lineage: approximately 80% of SNPs with allelic associations to chromatin accessibility in human NPCs or neurons only carry such an association in one of these two cell types (54). Thus, experimental study of putative disease-associated variants requires firm hypotheses on where (cell type), when (development/differentiation), and how the regulators are expressed/active and biologically relevant. For example, a hypothetical study of regulatory variants affecting action of a candidate ASD gene like *CHD8*—expressed most highly in early development (55)—using mature neurons may not detect any effects because, while matched on cellular context, the temporal context is inappropriate for examining the candidate gene. The importance of cell type in identifying regulatory variants is explicitly highlighted by the enrichment of GWAS SNPs with tissue-specific associations to gene expression (56,57). **MPRAs are thus poised to be applied to study of cell type-specificity of regulatory variation both *in vitro* and, as we will demonstrate, *in vivo*.**

Work toward assessing and predicting the effects of biological context on genomic variation has been recently illustrated in computational genomics. The Critical Assessment of Genome Interpretation 5 (CAGI5) consortium performed an MPRA of saturation-mutagenized human regulatory elements and disease-associated promoters in numerous cell types. A subset of these data were provided to analysts, who were then challenged to computationally predict functionality and effect sizes of held-out variants (58). The end project focused on identifying the most informative datasets (*e.g.*, chromatin modification data from ENCODE) for cell-type specific regulatory variant prediction. A key finding was that the most informative datasets often came from the same or similar cell types as that for which predictions were made.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Complimentarily, MPRAs focused on neuropsychiatric disorder associated variation stand to benefit *from* high-information datasets by aiding variant prioritization for assay inclusion. Several datasets on synthetic UTRs (27,29), RNA binding proteins (59,60), and postmortem human brain multi-omics (61–70) have become available in recent years. Integrative computational analyses have brought these datasets together predict functional variation in SCZ, bipolar disorder, and ASDs (71,72); however, these predictions have not yet been systematically tested. These constitute high-priority candidates for experimental validation by MPRA. Finally, inclusion of MPRA data with regulatory-epigenomic annotations improves machine learning predictions of functional variants (73). As such, MPRAs and omics form a symbiotic loop whereby MPRA results refine regulatory omics data, improving variant prioritization for future computational and functional investigations.

In spite of their potential, few MPRAs have examined regulatory variation while considering both cell type and –omic predictions. Tewhey, et. al (41) implemented an MPRA in human lymphoblastoid cell lines (LCLs) to identify functional regulatory variants among nearly 30,000 eQTL SNPs from LCLs, thus maintaining an appropriate cellular context for testing the discovered variants. Over 3,400 active regulatory sequences were identified, including 850 activity-modulating variants (24%), detectable even though effect sizes generally were under 2-fold. More recently, Choi, et. al (74) prioritized over 800 SNPs—guided by fine mapping and epigenomic signals—from 16 melanoma GWAS loci, then assayed the variants for transcriptional-regulatory activity in melanocytes *in vitro*. Candidate variants with concordant eQTL signal in independent melanocyte data were prioritized for follow-up, ultimately enabling their experimental demonstration of biophysical (TF binding), molecular (target gene expression), cellular (growth rate), and *in vivo* (melanoma rate in target over-expressing zebrafish) variant mechanisms. These experiments exquisitely illustrate MPRA's capacity for sensitivity, context specificity, and high discovery rates, especially when integrating both association data and multi-omic annotations to nominate variants.

However, the ability of MPRAs to address clinically-relevant questions has, to date, largely been restricted by *in vitro* implementations, which can neither model the complex cell-type interactions of tissues—especially those of the brain—nor can they model tissues' full breadth of sex/developmental, environmental, or pharmacological complexity **(Figure 2)**. Critically, even the most complex *in vitro* models of brain tissue—organoids—lack almost 50% of cell type-specific accessible chromatin regions found in their *in vivo* counterpart (fetal brain) (75), suggesting that *in vitro* MPRAs alone cannot dissect functional variation in neuropsychiatry. As preclinical research tools, *in vivo* MPRAs would greatly expand the ability to assess regulatory activity in native cellular and physiologic contexts. Two *in vivo* MPRA studies to date from Shen, et. al, confirm that MPRA is plausible for the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

brain (mouse cortex) *in vivo.* In the first, a library 45,000 barcodes corresponding to human brain open chromatin regions were packaged in adeno-associated virus 9 (AAV9) and transduced into adult mouse brain by stereotaxy (45). Consistent with open chromatin as a proxy for transcriptional regulation, many of the strongest accessibility peaks also drove gene expression in the mouse brain. In the second study, a single candidate SNP from a locus associated with bipolar disorder was prioritized by epigenomic data and ultimately demonstrated to have functional consequences; the two alleles of this sequence region were paired to 20 barcodes each, delivered by electroporation of embryonic mouse brains, which were explanted, maintained *ex vivo* for two days, and then assayed (44). While the two studies represent a major step toward assessing gene regulation in intact tissues, the barcode readout nonetheless comes from 'bulk' tissue RNA, leaving a limitation (albeit acknowledged) regarding cell-type effects on regulation.

The frontiers of context-driven MPRA: A Proof of Principle *In Vivo*, Cell-Type Specific MPRA

To determine whether MPRA can be used to assay cell-type specific gene regulation in the *in vivo* setting of the brain, we generated four barcoded sets of AAV-compatible plasmids, containing a human promoter from pan-cellular (*PGK2)* (76)*,* excitatory-neuronal (*CAMK2A),* or astrocytic (*GFAP*) (77) genes, or a minimal promoter (mouse *Hsp68*) (34) alone.

First, *CAMK2A*, *GFAP*, and minimal-promoter-only (*Hsp68)* viruses were individually injected into P1 mice to give widespread transduction of the brain (78). Brains were collected and sliced at P21 to examine cell-type expression patterns of dsRed under each promoter, confirming the neuronal, astrocytic, and non-specificity of these three promoters, respectively (**Figure 3**).

Subsequently, the viruses were mixed at a titer of 1:1:1:1 and injected into the P1 brain of SNAP25-TRAP mice (n=2), from which neuronal RNAs can be selectively enriched using the TRAP method (79,80). At P28, brains were collected for TRAP to obtain neuronal and brain-wide RNAs for barcode sequencing alongside DNA from the viral mix delivered.

As expected, the neuronal TRAP fractions relative to the whole brain homogenate displayed increased RNA/DNA ratio (*i.e.*, expression) of the *CAMK2A*-driven barcodes, and relative depletion of the astrocytic *GFAP*-driven barcodes (**Figure 4F**), indicating that we were able to successfully identify cell-type differential activity of these regulators. The calculated expression on a per-barcode basis was highly replicable between the two animals ($R^2 > 0.99$) (**Figure 4D and 4E**). In all, this experiment demonstrates that MPRAs are not restricted to tissue culture models and have the capacity to be

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

executed in specific cell types in model organisms, enabling functional assessment of gene regulation in higher-fidelity developmental, hormonal, and physiologic contexts.

Complementary methods in high-throughput study of DNA and RNA regulatory elements

A primary limitation of MPRA is the inability to test regulators in their endogenous genomic position and sequence context. This prevents identification of target genes (for transcriptional regulators or their variants) and confirmation of a variant or element's relevance in its native genomic context. Sequence-specific targeting methods using CRISPR/Cas9 have enabled several additional techniques for probing molecular and cellular effects of regulatory variation, with the caveat that, in contrast to MPRA, these techniques do not currently allow for the multiplexed study of post-transcriptional/translational regulators. Nonetheless, these techniques complement MPRA well in that they enable study of the host genome directly. *Perturb-seq* (81) combines genewise perturbation by CRISPR with single-cell RNA-seq to identify gene sets dysregulated by loss of function of each candidate gene. These have, for example, been used to discover co-transcribed gene networks involved in neuronal remodeling (82), and for *in vivo* assessment of a set of genes discovered to harbor *de novo* loss of function mutations in ASDs (83). CRISPR editing has been used *in vitro* to assess single transcript variant effects by comparing reference and allelic RNA and genomic DNA abundances in edited cultures(84). While this approach currently requires a separate culture of cells for each assayed variant, it still enables practicable validation/follow-up of MPRA candidates by the dozens using, *e.g.,* 96-well plates. To our knowledge, such assays have not been conducted at a genome-wide scale in psychiatric disease, but have been used to identify genes that alter expression of the Parkinson's-associated *PARKIN* (85).

Cis-regulatory MPRAs do not disentangle target gene(s) from functional regulators. Fortunately, CRISPR-derived methods using a mutagenically-'dead' Cas9 (dCas9) conjugated to a transcriptional activator or repressor allow an experimenter to target and potentiate or repress endogenous genomic regulatory elements (CRISPRa and CRISPRi, respectively) to assess altered gene expression and other outcomes. These technologies are already online in state-of-the-art human neuroscience models: a recent CRISPRi study knocked down over 2000 genes by targeting their promoters in iPSC-derived excitatory neurons, defining gene roles in their survival, differentiation, and proliferation. Like MPRAs, such assays retain the ability to assess cellular context—in this case, a subset of cells were co-cultured with astrocytes, revealing that knockdown of certain genes had different consequences on neuronal survival in co-culture conditions (86). The cell-type specificity of CRISPRi has also been leveraged to study ASD-associated gene knockdown effects in an etiologically relevant

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cell type, NPCs (87). A recently introduced extension of CRISPRi targets intergenic regulators along with fluorescent *in situ* hybridization against genes in the same chromodomain to identify regulator-target pairs. Fluorescence-intensity sorting into bins and subsequent RNA sequencing of stained cells then could be used to pair regulators (via guide RNA sequence) and target genes (altered fluorescent intensity associated to a guide RNA's presence) (88). While this assay was performed in K562 cells, it is not hard to envision its extension to neural cell types *in vitro* or *in vivo*. Altogether, CRISPR-based follow-up of MPRA candidates to define target genes and verify of genomic activity of regulators/variants will be key to developing insights in psychiatric genomics.

Part 2: MPRAs as an avenue to dissect multiallelic and polygenic mechanisms of neuropsychiatric traits

While MPRAs cannot intrinsically scale up to functional demonstration of cell-, tissue-, or behavior-level phenotypes, they have the potential to provide key information to guide molecular hypotheses for how these higher-order phenotypes emerge from large sets of regulators and/or their target genes. We will divide this discussion into a brief examination of *cis* variation—that is, the study of multiple elements in the same linkage block —and examination of variants in *trans* space—that is, defining shared and recurrent features among MPRA-nominated functional variants across the genome that may collectively underlie large portions of polygenic disease risk.

The utility of MPRAs for parsing linked variation

One challenge in parsing loci implicated in common variant association studies is that multiple variants in the region are equally statistically associated, and thus equally plausible functional mediators of risk. Indeed, in statistical genetics, repeating a GWAS or eQTL association analysis after conditioning on the lead SNP within a block often reveals one or more additional independent variants associated with traits or gene expression, respectively (*e.g.*, (70,89,90)); in fact, nearly half (~8,000) of brain expressed genes have ≥1 conditional eQTL (91)Reporter assays have been used to systematically evaluate such linked sets of variants. One disease-oriented, small scale example used epigenomic data to identify 16 enhancers spread over a few hundred kilobases near the *Ret* gene, known to be downregulated in Hirschsprung's disease (failure of terminal colonic nerves to form *in utero*). The 16 putative regulators were assayed for allele-differential activity in neuroblastomas to model regulatory activity in the disease-relevant cell type, neural crest cells. These were likewise assayed with transient expression in mouse embryos driving a LacZ reporter to identify the crest-relevant regulatory sequences. They also validated roles for regulator-binding TFs via siRNA knockdown. In all, this identified three functional SNPs in linkage disequilibrium(92)**.** While this was a

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

small scale example, it is easy to imagine how MPRAs might be used to dissect the large and numerous blocks of common variation associated with psychiatric disease (3,4).

While the *RET* SNPs were several hundred kilobases apart, another challenge is with variants in very close proximity, potentially within the same regulatory element. While oligo synthesis is limited in length, MPRA-based assay of such variants spanning up to 700bp is now possible with use of PCR amplification of oligos with uniquely complimentary ends (93). Likewise, 'capture-and-clone' MPRA designs, often using STARR-seq architecture to simplify cloning, fragment genomic DNA or capture e.g. ChIP-seq DNA fragments to test larger fragments than attainable via oligonucleotide synthesis(45,94–96)**.** Altogether, MPRAs allow for discovery of multiple functional variants per linkage region, as well as close-proximity discovery and disentanglement of multi-variant regulatory effects.

It is worth mentioning that MPRA could be of utility in the longer term as a technical tool to improve GWAS methods. Unusual linkage patterns in genomic regions such as the major histocompatibility complex confound localization of trait associations, leaving the region excluded from analyses or nebulously associated in others—over 200 traits (64), with MDD among them (65). Comprehensive MPRA of variants from this region across a broad panel of cell types would enable identification of cell/tissue-specific functional variants, which could then be placed on SNP genotyping arrays, partially sidestepping linkage limitations in understanding this region by collecting genotype information *at* the functional variant positions. Such efforts at a consortium scale could likewise characterize functionality and properties of broader sets of both untranscribed and untranslated variants in the same manner in each of several cell types.

<u>The utility of MPRAs in identifying commonalities from variants across the genome</u>

The most vexing question that remains after individual functional variant mechanisms are elucidated is how variants *collectively* contribute to phenotypic risk. MPRAs provide several ways to begin addressing this question: 1) by identifying shared regulatory features across several functional risk variants; 2) identifying functional modules enriched for genes dysregulated by disease-associated variants; 3) by providing functional annotations to variants that can be integrated into computational genomic approaches; and finally, 4) by enabling study of variant-by-environment interactions using the same MPRA libraries across conditions.

Firstly, MPRA experiments running the gamut from basic regulatory genomics to human traits and variation have defined scoping 'regulatory grammars' of assayed contexts. Identification of functional

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variants in the MPRA setting enables similar establishment of the "regulatory grammar" of a trait or disease. Functional variants identified by MPRA across several UTRs may feature a specific RBP's binding site, for example. Likewise, variants associated with a trait could be more likely to fall in particular TF binding sites or be enriched cell-type specific marks of genomic regulation. Evidence of this convergence is seen in *de novo* variants associated with ASD: several distinct variants disrupt binding sites for a single TF, *NFIX* (97). Such an approach was taken in a recent MPRA of lupus-associated SNPs, where identified functional variants were intersected with TF ChIP-seq datasets from pertinent cell types (leukocytes), identifying sets of recurrently disrupted TF binding sites (98). Assays of downstream consequences of variation also confirm biological convergence across association loci. A four-gene-target CRISPRi/a assay revealed that schizophrenia risk genes act synergistically via shared influence on synaptic activity, and concurrent alteration of expression of all four genes results in a cellular transcriptome more accurately reflective of postmortem schizophrenia brain tissue(99). For both rare and common variants, identifying common regulators among risk genes provides information which can refine predictions of disease-related cell types based on their expression of such TFs, RBPs, or epigenomic marks.

Secondly, genes and gene networks affected by statistically associated variation are often predicted using MAGMA (100), which in essence scores genes based on proximity to an associated variant and its linkage partners. Resulting gene sets are subjected to analyses such as Gene Ontology enrichment or are examined for enrichment in WGCNA (co-transcriptional) networks from candidate tissue types to identify pathways and mechanisms on which these genes converge. While its use is ubiquitous in genomic studies, standard MAGMA gene association statistics for psychiatric disorders only modestly correlate to those from a tissue-specific, chromatin configuration-aware modification of MAGMA (101), suggesting that biological hypotheses from MAGMA gene sets may miss disease-associated genes in brain. Being able to refine implicated genes by functional validation using—or in follow-up to—MPRA will help to benchmark such approaches and refine prediction convergence with 'truly' dysregulated candidate genes.

Thirdly, epigenomic data alone is not comprehensively predictive of active regulators. However, well-informed analyses of human genetic findings rely heavily on such annotations to convert associations into biological hypotheses. Critically, these epigenomic data—unlike MPRA data—can be collected from postmortem human tissue. However, a symbiotic loop is possible wherein verified regulators implicated by MPRA could be used to train and improve annotation of epigenomic algorithms —which in turn could be assessed by MPRA and used as another training set. Refined interpretation of epigenomic data and variant-expression association can then improve myriad analyses, such as

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variant enrichment in classes of genomic features, as well as disease gene identification. Other high-order analyses, such as TWAS (102) and Predixcan (103) intersect gene expression QTLs (eQTL) with trait-associated variants to predict expression differences between cases and controls, thus identifying dysregulated gene sets. MPRA data can disentangle which eQTL SNPs are truly functional from those associated only due to LD. Layering in MPRA data as weights or qualifiers of SNPs included in these analyses ought to refine disease-gene set associations much as discussed above. Altogether, MPRA can serve to refine both epigenomic and genic definitions of truly causal disease features.

Finally, the context-specificity of MPRA (**Fig 2**) represents a newfound ability to assess variant effects on gene regulation *en masse* under different biological and environmental contexts, including with *in vivo* models. While issues of convergent disease effects across genes and regulators are indeed complex, environmental effects—perhaps most canonically, stress—on these regulators are questions at the forefront of understanding polygenic risk in neuropsychiatric disorders. Pharmacologic variables have been successfully tested in MPRA, namely in the identification of glucocorticoid-responsive (104) and p53-responsive (96) regulatory elements. MPRAs could further be layered with concurrent gene perturbations (*e.g.*, knockdown of a putative regulator) or cell culture conditions for *in vitro* identification of variant-environment interactions. However, the most exciting function of MPRAs may be the opportunity to study disease-associated factors that cannot be (well) modeled *in vitro*, such as sex or stress, by moving these assays *in vivo*. Such studies can clarify sets of variants which depend on specific exogenous (*e.g.*, stress) or endogenous (*e.g.*, sex) contexts. Such variants could indicate molecular mechanisms behind conditional disease risk in a disorder such as depression.

Conclusion

MPRA presents unique opportunities to dissect polygenicity of psychiatric disorders via simultaneous identification of functional variants across identified risk space. Beyond the primary benefits of identifying ‘true positive’ functional variants in specific biological and environmental contexts, MPRAs stand to rapidly broaden, deepen, and refine hypotheses and mechanisms of both noncoding disease risk and of gene-regulatory architecture itself.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1 | **Cloning, Delivery, and Analysis Approaches in MPRA. A)** Visual key for subsequent panels. **B)** Enhancer MPRAs most commonly clone a pool of custom oligonucleotide pools containing sequences to be assayed, each paired to multiple unique barcode sequences, into a vector. This plasmid pool is collected and a promoter-reporter cassette is subsequently inserted between the elements and barcodes, such that the element of interest is upstream of a promoter and the barcode is in a 3'UTR. In the case of STARR-seq (not shown), the paradigm in panel **D** is instead used, with the *cis*-regulatory element acting as its own barcode. **C)** 5'UTR assays likewise use a two-step cloning assay, placing elements immediately *down*stream of a promoter. A reporter gene itself is inserted between the element and barcode. **D)** 3'UTR assays place both the elements of interests, immediately adjacent to barcodes, downstream of a promoter-reporter in a single cloning step. **E)** Sequence pools from approaches B-D can then be packaged into AAV or lentivirus (if in a compatible vector), or used for assay as plasmid directly. **F)** The plasmid or viral MPRA library can be delivered to cells in culture or in vivo tissue. After transfection/transduction, nucleic material is collected to generate sequencing libraries to quantify expression of the delivered elements. **G)** MPRA analysis centers on taking the ratio of RNA/DNA counts (or counts per million), represented by the sequence fragments at top left, as a measure of expression.

Figure 2 | **Regulatory assays are influenced by a range of conditions, from environment to sequence context**. The range of conditions that influence regulatory assays (from top to bottom) starts when considering the environment, *e.g.*, sex, time, and pharmacology. These parameters have the potential to affect various –omic profiles in a given system. The next level of consideration is the organism, which can include human-derived tissue or one of the many model organisms. Human genomic context is ideal for studying the biology of human disease – though a comparatively limited scope of techniques for human-derived tissues exists. Next, one should consider the selected cell type(s) and whether to assay *in vitro* or *in vivo*. Each of these provides a unique set of benefits, and one approach can be used to verify findings from the other as a form of validation (45,74). In the case of modeling the brain and psychiatric genetic variants, cell type-specific/enriched MPRAs *in vivo* constitute the highest-fidelity model of variant effects by accounting for regulatory effects of endogenously interacting cell types. Lastly, the sequence context will be influenced by the delivery method, which results in transcription from extragenomic or intragenomic MPRA DNAs. In either case, only limited length of sequence surrounding a feature of interest is preserved (*e.g.*, in ~120bp tiles of genomic sequence in custom oligonucleotide cloning, or ≤ 1kb in clone-and-capture methods), preventing assessment of any interactive effects from elements further away. (A recent study suggests that size of a tile negatively correlates with reproducibility of expression driven compared to that driven by ~120bp tiles, emphasizing the importance of this consideration (93)). While AAV-

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transduced episomes gain histones (105) and chromosome-like nucleosome spacing (106), it is unknown whether gene-regulatory histone marks on these episomes mirror those of endogenous regulatory chromatin. For these reasons of both local sequence context and chromatin context, we suggest corroboration of MPRA findings in native genomic settings, by, for example, introducing the variant to the genome of a cell line using CRISPR methods.

Figure 3 | **Immunofluorescent confirmation of cell type-specificity of promoters delivered by AAV9.** All images are from mouse brain transduced with AAV9 at P2 and collected at P27. The AAV delivered contained dsRed under the control of the hsp68 minimal promoter, with or without an additional promoter upstream. Astrocytes were labeled with anti-GFAP and neurons with anti-NeuN antibodies. White arrowheads indicate examples of transduced neurons; white arrows indicate example transduced astrocytes. **A)** *Hsp68* alone drives dsRed expression predominantly in neurons in both the cortex (top row) and hippocampal granule layer (bottom row). Only three astrocytes (two shown) were found expressing dsRed across three examined slices of transduced forebrain. **B)** The *CAMK2A* promoter exclusively drove neuronal expression, shown in the hippocampal dentate gyrus granule layer. **C)** The GFABC1D promoter (subregion of the *GFAP* promoter) drove predominantly astrocytic dsRed expression, shown near the granule cell layer of the hippocampus. Expression is predominantly in astrocytes, such as those ensheathing blood vessels as shown, with rare, comparatively low expression in neurons.

Figure 4 | **Design, replicability, and cell type-specificity of a small-scale, proof of principle *in vivo* MPRA. A-C)** Design of the MPRA plasmids. For each promoter, multiple barcodes were PCR coupled to each promoter (**A**), inserted into plasmid (**B**), and an hsp68-dsRed construct subsequently inserted (**C**). Plasmid pools were created subsequently packaged in AAV9 for each promoter separately. The final library consisted of 32 barcodes across the four promoters. **D-E)** Mouse brains transduced with an equal-titer mixture of all 4 promoter AAV9s and subjected to neuronal (SNAP25) TRAP showed strong replicability of barcode-wise expression in both the neuronal-enriched TRAP RNA fraction (**D**, R^2=0.9975) and total tissue (**E**, R^2=0.9927). **F)** Expression across barcodes in each replicate confirmed that the TRAP fraction was enriched for *Camk2a*-coupled barcodes (neuron-specific) and depleted for GFABC1D-coupled barcodes (astrocytic) as expected. ** $p \leq 5 \times 10^{-4}$, *** $p \leq 5 \times 10^{-8}$

References

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Consortium C-DG of the PG, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, *et al.* (2013): Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 45: 984–994.

2. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, *et al.* (2015): An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47: 1236–41.

3. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, *et al.* (2018): Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 50: 668–681.

4. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, *et al.* (2014): Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511: 421–7.

5. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, *et al.* (2020): Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat Neurosci* 1–9.

6. Ryan NM, Lihm J, Kramer M, McCarthy S, Morris SW, Arnau-Soler A, *et al.* (2018): DNA sequence-level analyses reveal potential phenotypic modifiers in a large family with psychiatric disorders. *Mol Psychiatr* 23: 2254–2265.

7. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, *et al.* (2014): Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95: 535–52.

8. An J-Y, Lin K, Zhu L, Werling DM, Dong S, Brand H, *et al.* (2018): Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* 362: eaat6576.

9. Gamazon ER, Segrè AV, Bunt M van de, Wen X, Xi HS, Hormozdiari F, *et al.* (2018): Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat Genet* 50: 956–967.

10. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012): Linking disease associations with regulatory information in the human genome. *Genome Res* 22: 1748–1759.

11. Benton ML, Talipineni SC, Kostka D, Capra JA (2019): Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *Bmc Genomics* 20: 511.

12. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, *et al.* (2015): Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317–30.

13. Brown AA, Viñuela A, Delaneau O, Spector TD, Small KS, Dermitzakis ET (2017): Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat Genet* 49: 1747–1751.

14. Mao F, Wang L, Zhao X, Li Z, Xiao L, Rao RC, *et al.* (2019): Post-transcriptionally impaired de novo mutations contribute to the genetic etiology of four neuropsychiatric disorders. *Biorxiv* 175844.

15. Turner TN, Eichler EE (2018): The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. *Trends Neurosci* 42: 115–127.

16. Mayr C (2017): Regulation by 3′-Untranslated Regions. *Annu Rev Genet* 51: 171–194.

17. Price AJ, Hwang T, Tao R, Burke EE, Rajpurohit A, Shin JH, *et al.* (2019): Characterizing the nuclear and cytoplasmic transcriptomes in developing and mature human cortex uncovers new insight into psychiatric disease gene regulation. *Genome Res* 30: 1–11.

18. Shadrin AA, Frei O, Smeland OB, Bettella F, O'Connell KS, Gani O, *et al.* (2019): Annotation-Informed Causal Mixture Modeling (AI-MiXeR) reveals phenotype-specific differences in polygenicity and effect size distribution across functional annotation categories. *Biorxiv* 772202.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J (2009): High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* 27: 1173–1175.

20. Kinney JB, Murugan A, Callan CG, Cox EC (2010): Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *P Natl Acad Sci Usa* 107: 9158–63.

21. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G (2015): Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* 163: 698–711.

22. Wong MS, Kinney JB, Krainer AR (2018): Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. *Mol Cell* 71: 1012-1026.e3.

23. Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B (2019): Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* 176: 549-563.e23.

24. Safra M, Nir R, Farouq D, Slutzkin IV, Schwartz S (2017): TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res* 27: 393–406.

25. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, *et al.* (2018): Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* 50: 874–882.

26. Rieger MA, King DM, Cohen BA, Dougherty JD (2018): CLIP-Seq and massively parallel functional analysis of the CELF6 RNA binding protein reveals a role in destabilizing synaptic gene mRNAs through interaction with 3'UTR elements in vivo. *Biorxiv* 401604.

27. Cottrell KA, Chaudhari HG, Cohen BA, Djuranovic S (2018): PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. *Nat Commun* 9: 301.

28. Rabani M, Pieper L, Chew G-L, Schier AF (2018): A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. *Mol Cell* 70: 565.

29. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G (2019): Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* 37: 803–809.

30. Litterman AJ, Kageyama R, Tonqueze OL, Zhao W, Gagnon JD, Goodarzi H, *et al.* (2019): A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res* 29: 896–906.

31. King DM, Maricque BB, Cohen BA (2018): Synthetic and genomic regulatory elements reveal aspects of cis regulatory grammar in Mouse Embryonic Stem Cells. *Biorxiv* 398107.

32. Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, *et al.* (2017): Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc National Acad Sci* 114: E1291–E1300.

33. Levo M, Avnit-Sagi T, Lotan-Pompan M, Kalma Y, Weinberger A, Yakhini Z, Segal E (2017): Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Mol Cell* 65: 604-617.e6.

34. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA (2014): High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* 24: 1595–1602.

35. Fiore C, Cohen BA (2016): Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res* 26: 778–86.

36. Gisselbrecht SS, Barrera LA, Porsch M, Aboukhalil A, Estep PW, Vedenko A, *et al.* (2013): Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nat Methods* 10: 774–780.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37. White MA, Myers CA, Corbo JC, Cohen BA (2013): Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc National Acad Sci* 110: 11952–11957.

38. Maricque BB, Chaudhari HG, Cohen BA (2018): A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol* 37: 90–95.

39. Klein JC, Keith A, Rice SJ, Shepherd C, Agarwal V, Loughlin J, Shendure J (2018): Functional Testing of Thousands of Osteoarthritis-Associated Variants for Regulatory Activity. *Biorxiv* 379727.

40. Castaldi PJ, Guo F, Qiao D, Du F, Naing ZZC, Li Y, *et al.* (2019): Identification of Functional Variants in the FAM13A Chronic Obstructive Pulmonary Disease Genome-Wide Association Study Locus by Massively Parallel Reporter Assays. *Am J Resp Crit Care* 199: 52–61.

41. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, *et al.* (2016): Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* 165: 1519–1529.

42. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, *et al.* (2016): Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* 165: 1530–1545.

43. Myint L, Wang R, Boukas L, Hansen KD, Goff LA, Avramopoulos D (2018): A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential. *Biorxiv* 447557.

44. Shen SQ, Kim-Han JS, Cheng L, Xu D, Gokcumen O, Hughes AEO, *et al.* (2019): A candidate causal variant underlying both higher intelligence and increased risk of bipolar disorder. *Biorxiv* 580258.

45. Shen SQ, Myers CA, Hughes AEO, Byrne LC, Flannery JG, Corbo JC (2016): Massively parallel cis -regulatory analysis in the mammalian central nervous system. *Genome Res* 26: 238–255.

46. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A (2013): Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* 339: 1074–1077.

47. Mogno I, Kwasnieski JC, Cohen BA (2013): Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* 23: 1908–1915.

48. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA (2012): Complex effects of nucleotide variants in a mammalian cis-regulatory element. *P Natl Acad Sci Usa* 109: 19498–503.

49. Maricque BB, Dougherty JD, Cohen BA (2016): A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis -regulatory activity in neural cells. *Nucleic Acids Res* 45: gkw942.

50. Wagnon JL, Briese M, Sun W, Mahaffey CL, Curk T, Rot G, *et al.* (2012): CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *Plos Genet* 8: e1003067.

51. Dougherty JD, Maloney SE, Wozniak DF, Rieger MA, Sonnenblick L, Coppola G, *et al.* (2013): The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. *J Neurosci* 33: 2732–2753.

52. Jia L, Mao Y, Ji Q, Dersh D, Yewdell JW, Qian S-B (2020): Decoding mRNA translatability and stability from 5′UTR. *bioRxiv* 2020.03.13.990887.

53. Inoue F, Kreimer A, Ashuach T, Ahituv N, Yosef N (2019): Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*. https://doi.org/10.1016/j.stem.2019.09.010

54. Liang D, Elwell AL, Aygün N, Lafferty MJ, Krupa O, Cheek KE, *et al.* (2020): Cell-type specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Biorxiv* 2020.01.13.904862.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

55. Bernier R, Golzio C, Xiong B, Stessman HA, Coe BP, Penn O, *et al.* (2014): Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 158: 263–76.

56. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, *et al.* (2017): Genetic effects on gene expression across human tissues. *Nature* 550: 204–213.

57. Kopp N, Climer S, Dougherty JD (2015): Moving from capstones toward cornerstones: successes and challenges in applying systems biology to identify mechanisms of autism spectrum disorders. *Frontiers Genetics* 6: 301.

58. Shigaki D, Adato O, Adhikar AN, Dong S, Hawkins-Hooker A, Inoue F, *et al.* (2019): Integration of Multiple Epigenomic Marks Improves Prediction of Variant Impact in Saturation Mutagenesis Reporter Assay. *Hum Mutat* 40: 1280–1291.

59. Nostrand ELV, Pratt GA, Yee BA, Wheeler E, Blue SM, Mueller J, *et al.* (2019): Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Biorxiv* 807008.

60. Nostrand ELV, Freese P, Pratt GA, Wang X, Wei X, Blue SM, *et al.* (2018): A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *Biorxiv* 179648.

61. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, *et al.* (2018): Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362: eaat8464.

62. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, *et al.* (2018): Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Sci New York N Y* 362: eaat8127.

63. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, *et al.* (2012): An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

64. Moore JE, Pratt H, Purcaro M, Weng Z (2019): A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Biorxiv* 745844.

65. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, *et al.* (2016): Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 19: 1442–1453.

66. Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang Y-C, *et al.* (2019): CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data* 6: 180.

67. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Wucher V, Kasela S, *et al.* (2019): Cell type specific genetic regulation of gene expression across human tissues. *Biorxiv* 806117.

68. Consortium TGte, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, *et al.* (2015): The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348: 648–660.

69. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, *et al.* (2013): The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45: 580–5.

70. Consortium TB, Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, *et al.* (2018): Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* 21: 1117–1125.

71. Li M, Santpere G, Kawasawa YI, Evgrafov OV, Gulden FO, Pochareddy S, *et al.* (2018): Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Sci New York N Y* 362: eaat7615.

72. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, *et al.* (2019): Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* 51: 973–980.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

73. Li Y, Shi AH, Tewhey R, Sabeti PC, Ernst J, Kellis M (2017): Genome-wide regulatory model from MPRA data predicts functional regions, eQTLs, and GWAS hits. *Biorxiv* 110171.

74. Choi J, Zhang T, Vu A, Ablain J, Makowski MM, Colli LM, *et al.* (2019): Massively parallel reporter assays combined with cell-type specific eQTL informed multiple melanoma loci and identified a pleiotropic function of HIV-1 restriction gene, MX2 , in melanoma promotion. *Biorxiv* 625400.

75. Ziffra RS, Kim CN, Wilfert A, Turner TN, Haeussler M, Casella AM, *et al.* (2020): Single cell epigenomic atlas of the developing human brain and organoids. *Biorxiv* 2019.12.30.891549.

76. Li M, Husic N, Lin Y, Christensen H, Malik I, McIver S, *et al.* (2010): Optimal promoter usage for lentiviral vector-mediated transduction of cultured central nervous system cells. *J Neurosci Meth* 189: 56–64.

77. Lee Y, Messing A, Su M, Brenner M (2008): GFAP promoter elements required for region-specific and astrocyte-specific expression. *Glia* 56: 481–93.

78. Frisella WA, O'Connor LH, Vogler CA, Roberts M, Walkley S, Levy B, *et al.* (2001): Intracranial Injection of Recombinant Adeno-associated Virus Improves Cognitive Function in a Murine Model of Mucopolysaccharidosis Type VII. *Mol Ther* 3: 351–358.

79. Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, *et al.* (2008): A Translational Profiling Approach for the Molecular Characterization of CNS Cell Types. *Cell* 135: 738–748.

80. Dougherty JD, Fomchenko EI, Akuffo AA, Schmidt E, Helmy KY, Bazzoli E, *et al.* (2012): Candidate pathways for promoting differentiation or quiescence of oligodendrocyte progenitor-like cells in glioma. *Cancer Res* 72: 4856–68.

81. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, *et al.* (2016): Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167: 1853-1866.e17.

82. Alyagor I, Berkun V, Keren-Shaul H, Marmor-Kollet N, David E, Mayseless O, *et al.* (2018): Combining Developmental and Perturbation-Seq Uncovers Transcriptional Modules Orchestrating Neuronal Remodeling. *Dev Cell* 47: 38-52.e6.

83. Jin X, Simmons SK, Guo AX, Shetty AS, Ko M, Nguyen L, *et al.* (2019): In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with Autism risk genes. *Biorxiv* 791525.

84. Brandt M, Gokden A, Ziosi M, Lappalainen T (2019): A polyclonal allelic expression assay for detecting regulatory effects of transcript variants. *Biorxiv* 794081.

85. Potting C, Crochemore C, Moretti F, Nigsch F, Schmidt I, Manneville C, *et al.* (2017): Genome-wide CRISPR screen for PARKIN regulators reveals transcriptional repression as a determinant of mitophagy. *P Natl Acad Sci Usa* 115: E180–E189.

86. Tian R, Gachechiladze MA, Ludwig CH, Laurie MT, Hong JY, Nathaniel D, *et al.* (2019): CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron* 104: 239-255.e12.

87. Lalli MA, Avey D, Dougherty JD, Milbrandt J, Mitra RD (2019): Multiplexed single-cell autism modeling reveals convergent mechanisms altering neuronal differentiation. *Biorxiv* 862680.

88. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, *et al.* (2019): Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* 51: 1664–1669.

89. Consortium C-DG of the PG, Lee PH, Anttila V, Won H, Feng Y-CA, Rosenthal J, *et al.* (2019): Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* 179: 1469-1482.e11.

90. Deming Y, Filipello F, Cignarella F, Cantoni C, Hsu S, Mikesell R, *et al.* (2019): The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci Transl Med* 11: eaau2291.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.02.931337; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

91. Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, *et al.* (2018): Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. *Am J Hum Genet* 102: 1169–1184.

92. Chatterjee S, Kapoor A, Akiyama JA, Auer DR, Lee D, Gabriel S, *et al.* (2016): Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. *Cell* 167: 355-368.e10.

93. Klein J, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, *et al.* (2019): A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *Biorxiv* 2: 576405.

94. Arensbergen J van, Pagie L, FitzPatrick VD, Haas M de, Baltissen MP, Comoglio F, *et al.* (2019): High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* 51: 1160–1169.

95. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, *et al.* (2018): High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* 9: 5380.

96. Catizone AN, Uzunbas GK, Celadova P, Kuang S, Bose D, Sammons MA (2020): Locally acting transcription factors regulate p53-dependent cis-regulatory element activity. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkaa147

97. Amiri A, Coppola G, Scuderi S, Wu F, Roychowdhury T, Liu F, *et al.* (2018): Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Sci New York N Y* 362: eaat6720.

98. Lu X, Chen X, Forney C, Donmez O, Miller D, Parameswaran S, *et al.* (2020): Genome-wide discovery of SLE genetic risk variant allelic enhancer activity. *Biorxiv* 2020.01.20.906701.

99. Schrode N, Ho S-M, Yamamuro K, Dobbyn A, Huckins L, Matos MR, *et al.* (2019): Synergistic effects of common schizophrenia risk variants. *Nat Genet* 51: 1475–1485.

100. Leeuw CA de, Neale BM, Heskes T, Posthuma D (2016): The statistical properties of gene-set analysis. *Nat Rev Genet* 17: 353–364.

101. Sey NYA, Hu B, Mah W, Fauni H, McAfee JC, Rajarajan P, *et al.* (2020): A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci* 1–11.

102. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, *et al.* (2016): Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48: 245–52.

103. Consortium Gte, Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, *et al.* (2018): Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 9: 1825.

104. Vockley CM, D'Ippolito AM, McDowell IC, Majoros WH, Safi A, Song L, *et al.* (2016): Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell* 166: 1269-1281.e19.

105. Marcus-Sekura CJ, Carter BJ (1983): Chromatin-like structure of adeno-associated virus DNA in infected cells. *J Virol* 48: 79–87.

106. Penaud-Budloo M, Guiner CL, Nowrouzi A, Toromanoff A, Chérel Y, Chenuaud P, *et al.* (2008): Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. *J Virol* 82: 7875–85.

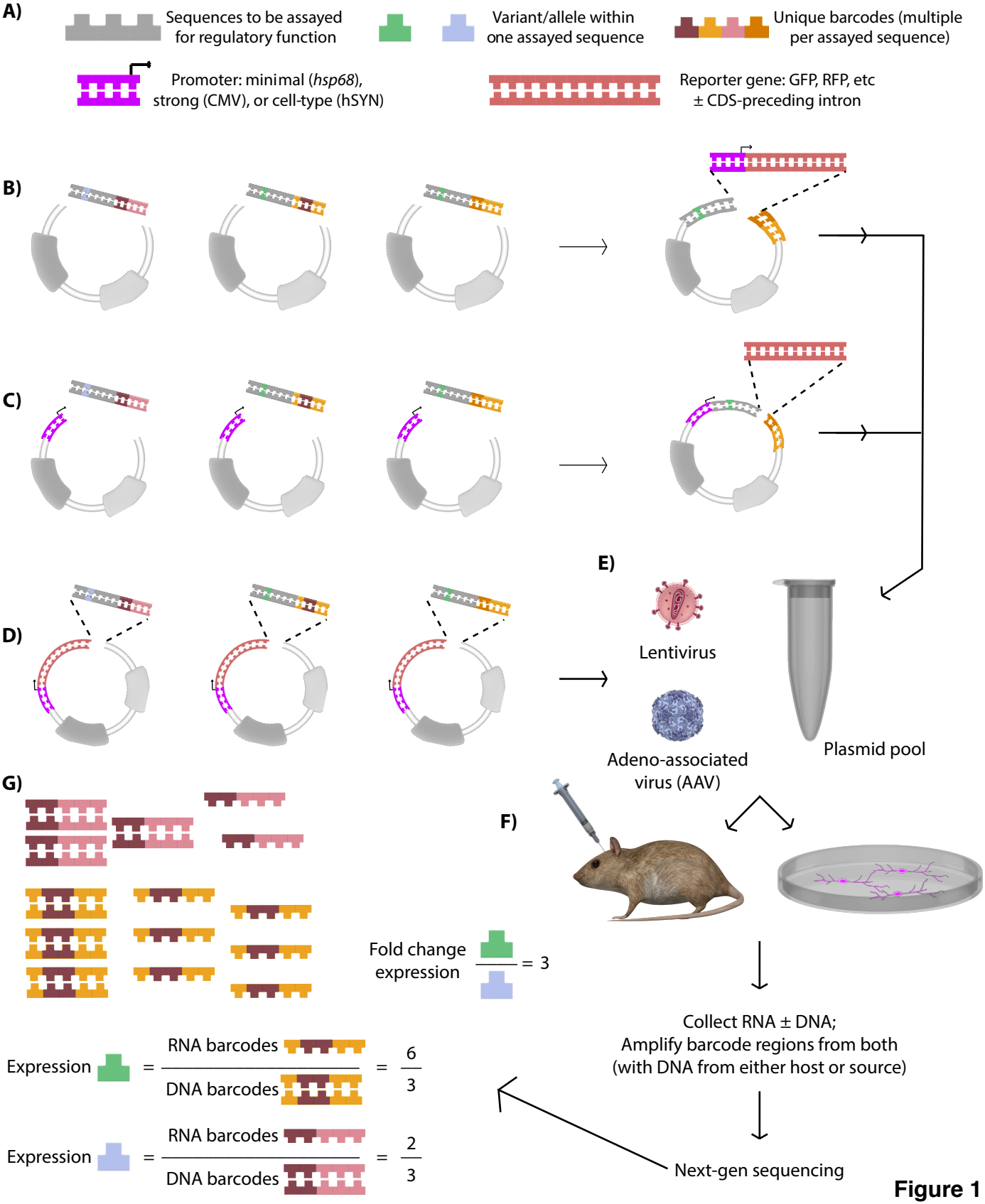
A)
Sequences to be assayed for regulatory function
Variant/allele within one assayed sequence
Unique barcodes (multiple per assayed sequence)
Promoter: minimal (hsp68), strong (CMV), or cell-type (hSYN)
Reporter gene: GFP, RFP, etc ± CDS-preceding intron
B)
C)
D)
E)
Lentivirus
Adeno-associated virus (AAV)
Plasmid pool
F)
G)
Fold change expression = 3
Expression = RNA barcodes / DNA barcodes = 6/3
Expression = RNA barcodes / DNA barcodes = 2/3
Collect RNA ± DNA;
Amplify barcode regions from both
(with DNA from either host or source)
Next-gen sequencing

Figure 1

Environment

Sex effects

Time and development

Pharmacology and Environment

Organism

Human

- Best context for studying the biology of human disease
- Several -omic datasets readily available

Model Organisms

- High-throughput
- Genetic similarity to human
- Several -omic datasets readily available

Cell Type

In vitro

- High-throughput, enabling comparisons across across multiple cell types
- Large range of applications
- Efficient/effective design

Multiple cell types can and should be assayed to identify disease-relevant cell types and regulatory elements

In vivo

- Whole-tissue complexity
- Clinically relevant
- Accurate molecular model system for sex and developmental variables

Sequence

Extragenomic (AAV, plasmid)

- High-throughput
- Efficient/effective design
- Practical *in vitro* AND *in vivo*
- Unknown extent of resemblance to endogenous chromatin
- Assays of 100bp-1kb of local sequence and contained elements

Variant of interest

Intragenomic (Lentivirus)

- Full host chromatin structure
- Time consuming and expensive
- Difficult to infect large number of cells *in vivo* via intracranial injection

Variant introduction to the genome at the endogenous variant position enables corroboration of MPRA findings in the presence of all cooperating regulatory elements, regardless of their distance from the variant

Figure 2

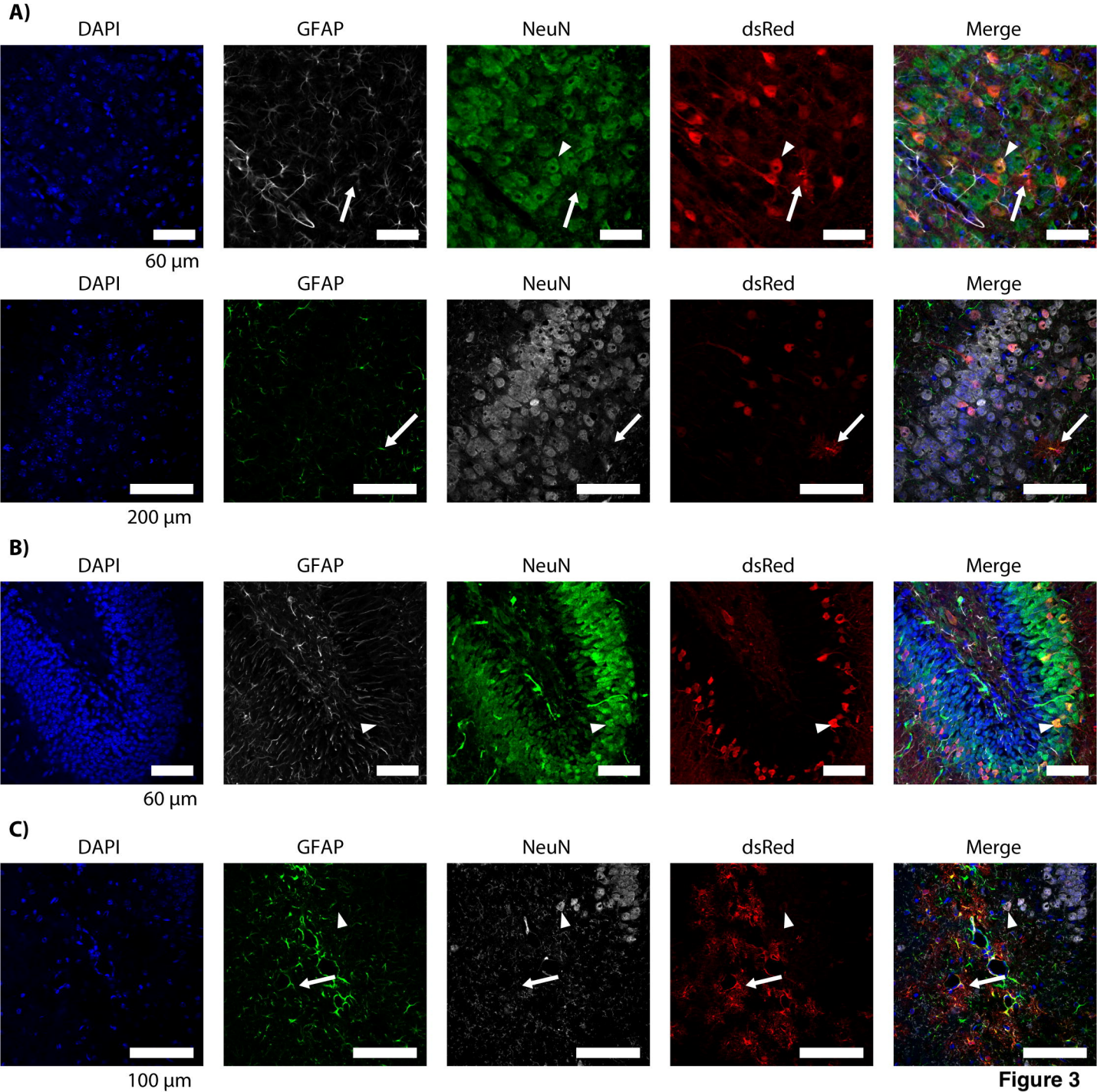
A)
DAPI
GFAP
NeuN
dsRed
Merge
60 µm
DAPI
GFAP
NeuN
dsRed
Merge
200 µm
B)
DAPI
GFAP
NeuN
dsRed
Merge
60 µm
C)
DAPI
GFAP
NeuN
dsRed
Merge
100 µm

Figure 3

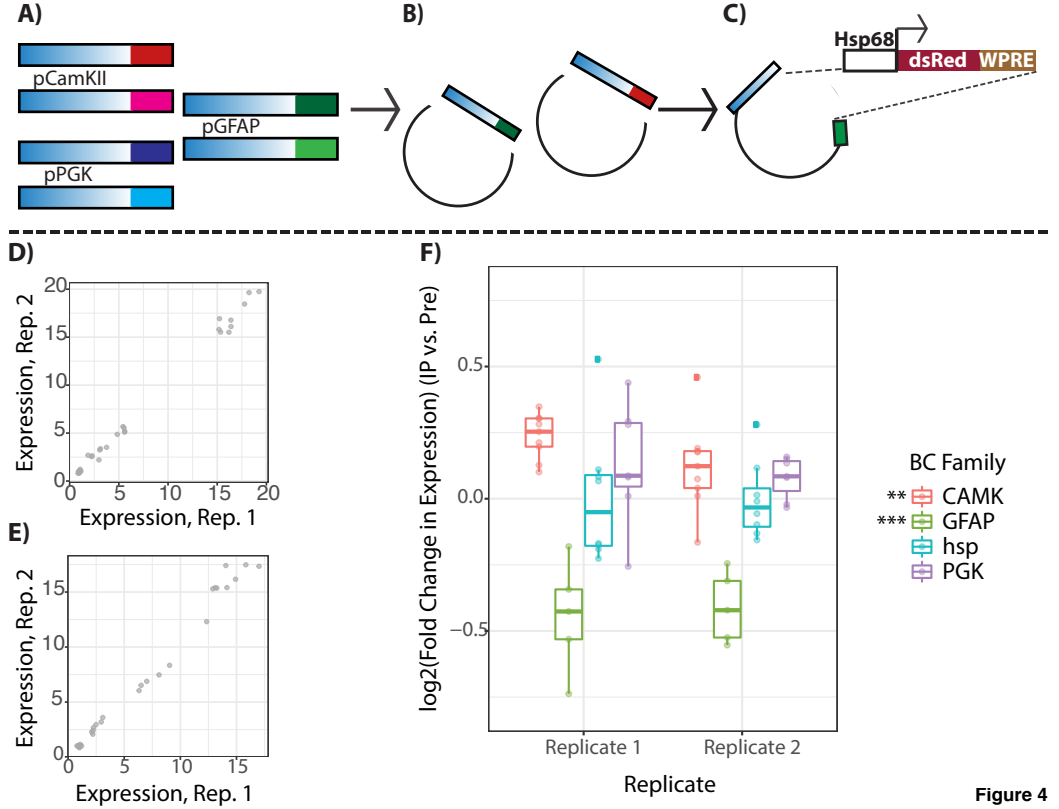
A)
pCamKII
pGFAP
pPGK
B)
C)
Hsp68
dsRed
WPRE
D)
Expression, Rep. 2
Expression, Rep. 1
E)
Expression, Rep. 2
Expression, Rep. 1
F)
log2(Fold Change in Expression) (IP vs. Pre)
Replicate 1
Replicate 2
Replicate
BC Family
** CAMK
*** GFAP
hsp
PGK

Figure 4