UC Davis

UC Davis Electronic Theses and Dissertations

Title

Defining Biological Impacts of Common GWAS Variants for Schizophrenia

Permalink

https://escholarship.org/uc/item/9dm356q0

Author

Warren, Tracy Leigh

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Defining Biological Impacts of Common GWAS Variants for Schizophrenia

By

TRACY LEIGH WARREN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Alex S. Nord, Chair

Jennifer L. Whistler

Karen Zito

Kyle D. Fink

Megan Y. Dennis

Committee in Charge

2024

Table of Contents

Preface iii
Acknowledgements iv
Abstract vi
Introduction 1
a. Preface 1
b. Mapping variants to biology in psychiatry 1
c. Contributions of this dissertation toward mapping GWAS variants onto psychiatric etiologies 14
Chapter 1: Development and deployment of *in vivo* MPRA to test enhancer activity in postnatal mouse brain 17
a. Preface 17
b. Resulting Publication 19
c. Appendix 56
Chapter 2: Schizophrenia-associated risk variants alter the function of regulatory sequences that show specific activity during adolescent brain development 60
a. Preface 60
b. MPRA protocol optimization 61
c. Pre-publication manuscript 68
Chapter 3: Identification of convergent pathways from common variants toward psychotic disorder endophenotypes 123
a. Preface 123
b. Resulting Publication 125
Concluding remarks 167

Preface

This dissertation describes three projects utilizing different genomic technologies to study the biological impact of common risk variants in schizophrenia. Experimental data described in Chapters 1 and 2 were generated in the lab of Dr. Alex Nord at UC Davis. Experimental data described in Chapter 3 were generated in the lab of Dr. Nord, as well as in the labs of Dr. Cameron Carter at UC Davis and Dr. Pak Chung Sham at the University of Hong Kong.

Chapter 1 describes methodologies to study gene regulation *in vivo* in mouse brain. These methods are subsequently utilized in Chapter 2 to test the impacts on gene regulation in mouse adolescent brain of schizophrenia risk variants identified in genome-wide association studies. Chapter 3 describes the cumulative effects of these common variants on symptomatology in humans with psychotic disorders, testing associations of mechanism-level polygenic risk scores for schizophrenia with endophenotypes of psychotic disorders in human subjects.

Acknowledgements

My PhD is the result of cumulative efforts of mentors, colleagues, collaborators, and funding agencies who have helped me along the way. I would sincerely like to thank the following people:

First and foremost, my graduate advisor and the principal investigator of my PhD lab, **Dr. Alex Nord**, for his incredible dedication to mentorship and my scientific and professional growth. Alex has, from the start, committed to me fully as his mentee. He has offered complete support when I've needed it, and absolute trust and independence when I was ready for it. I attribute so much of my development as a scientist to his fantastic mentorship skills and our excellent relationship during my PhD.

My dissertation committee **Drs. Jennifer Whistler, Karen Zito, Megan Dennis,** and **Kyle Fink,** as well as additional mentors in the UC Davis neuroscience community, **Drs. Kim McAllister** and **Gene Gurkoff**. Throughout my PhD, these mentors have pushed me to think critically about my research and consistently supported my endeavors. For all of their contributions to my scientific and professional growth, I offer them the deepest thanks.

Current and former members of the lab for their collaboration, support, mentorship, and social camaraderie. **Jason Lambert** and **Linda Su-Feher** were particularly vital in my initial training and the launching of many of my lab projects. **Cory Ardekani** and **Ana Santos** have been my mental support team, sharing in all of the successes and challenges that come with detangling *cis*-regulation in the genome. Our wonderful former undergraduate researcher **Mylena Corona** dedicated months to helping me extract and clean more than 300 samples of human DNA. **Ayanna Wade, Rinaldo Catta-Preta, Stephanie Lozano, Nick Chu, Emily Smith, Nic Seban, Ethan Fenton, Karol Cichewicz**, **Cesar Canales,** and **Yurong Wang** have offered invaluable

social and scientific support both in and out of the lab. Without our close and supportive lab community, I certainly would not have been able to achieve half the things I did during my PhD.

Drs. Cameron Carter, Pak Chung Sham, Tyler Lesh, and **Justin Tubbs**, who were close collaborators and offered strong support during my PhD. They worked hard to generate and help analyze much of the data described in Chapter 3, for which I am extremely grateful.

The **Floyd and Mary Schwall Fellowship in Medical Research**; the **Learning, Memory, and Plasticity T32**; the **Ruth L. Kirschstein Predoctoral Individual National Research Service Award F31**; the **Barbara Chapman Neuroscience Award**; and the **Ling-Lie Chau Graduate Student Award for Brain Research** have provided me with funds to support my research and training over the years. These awards have unlocked incredible opportunities during my career course that would likely not otherwise have been available to me, and I am immensely grateful for these chances.

Of course, my parents, **Don** and **Sallie**, and my sister, **Christi**, for not understanding what I do but excitedly talking about something that sounds genetics-y anyway. And finally, my cat, **Cat**, for reminding me that science can always wait until after breakfast, and my partner, **Ramiel Tamras**, for accepting scratches in my stead when I have ignored her.

Abstract

Schizophrenia and related psychotic disorders are highly polygenic. The vast majority of their associated risk variants lie in non-coding areas of the genome, making their biological impacts cryptic in nature. I describe in this dissertation the development and application of different genomic techniques for assigning biological relevance to known schizophrenia risk variants.

These risk variants are widely acknowledged to be enriched in regulatory elements in the genome, particularly in enhancers and silencers, which regulate gene expression in a cell type and temporally specific manner. In Chapter 1, I described a methodology I have helped drive in our lab to test for activity of hypothesized enhancers using AAV-based enhancer-reporter assays *in vivo* in postnatal mouse brain. In Chapter 2, I conducted this assay in juvenile and adolescent mice to test for regulatory activity at a prioritized set of noncoding schizophrenia risk SNPs. My results identified 39 active regulatory elements, 19 of which showed allelic differences in activity at the schizophrenia risk SNP. In Chapter 3, I collapsed both coding and noncoding common schizophrenia risk variants into their presumed targeted genes. I used gene ontology to divide these genes into *a priori*-selected hypothesized risk pathways. I subsequently tested for associations between these risk pathways and endophenotypes of psychotic disorders. Major findings showed associations between overall genome-wide risk and overall diagnosis of a psychotic disorder, as well as associations between glutamatergic and GABAergic polygenic risk and endophenotypes related to cognitive control and social functioning.

In my PhD, I used diverse methodologies to better characterize the mechanisms linking genomic schizophrenia risk variants to psychotic disorder symptoms. My findings highlighted the value of *in vivo* functional genomics screens, particularly for schizophrenia and other adolescent- and adult-onset disorders which are difficult to model using *in vitro* methods, and of using multiple levels of analysis to better define aberrant biology in psychiatric disorders.

Introduction

a. Preface

The overarching aim of my research is to improve mapping of common, GWAS-identified risk variants onto their disrupted biological pathways, toward the goal of identifying distinct etiologies in psychiatric disorders which may merit distinct treatments. Dividing patients by shared etiologies and applying treatments accordingly is typically described as *stratified medicine*[1]. To achieve stratified medicine in psychiatry, we will need a more precise understanding of the aberrant biology that underlies psychiatric symptoms. I have attempted to contribute to this improved understanding during my thesis work. To provide context for my research, I present my introduction in two parts. First, I present a review and commentary of the current state of biological pathway mapping in psychiatric genomics. I review relevant context in psychiatric research, major challenges in mapping genomic variants to psychiatric etiologies, and how genomic researchers are attempting to address these challenges. Second, I present a brief discussion of how my research aims to contribute to our understanding of how psychiatric risk variants impact biology. This introduction will frame my work into its clinical research context and highlight the need for identification of convergent etiologies in psychiatric disorders.

References

1. Schumann, G. *et al.* Stratified medicine for mental disorders. *European Neuropsychopharmacology* **24**, 5–50 (2014).

b. Mapping variants to biology in psychiatry

Psychiatry is unique as a subfield of medicine in that the disorders in its domain are almost exclusively syndromes rather than diseases[1,2]. That is, they are presently defined as clusters of correlated symptoms and behaviors with a relatively limited understanding of their underlying

biological causes. Many good reasons underlie this limitation of psychiatry. Critically, psychiatry is among the youngest medical disciplines, and focusing on the brain as the source of psychiatric phenomena only took root among clinicians and scientists in the 17th century[3] - a relatively recent revelation, considering that descriptions of the cardiovascular system and associated diseases have been documented from at least as early as the 16th century BCE[4]. Further, many psychiatric symptoms are, by nature, extremely difficult to test in the lab. While researchers have successfully developed animal behavioral assays for highly conserved psychological phenomena such as anxiety[5], social behavior[6], addiction[7], and memory[8], many psychiatric symptoms are difficult to model in traditional lab animals such as mice. This may be due to limited externally observable behaviors from modeled symptoms in lab animals, such as in the case of hallucinations and delusions[9], or even due to a dependence of the symptom on human-specific capabilities such as language[10]. Experimentation on the living human brain is ethically fraught, though technologies such as fMRI[11] and EEG[12] have made major breakthroughs in the study of live human brain activity without the application of invasive or unethical techniques.

As a result of this historical context and with good reason, psychiatry is behind its peers in medicine in the identification of diseases according to their biological roots and in the application of treatments in a targeted or stratified manner toward these roots. Consequently, many first-line psychiatric medications have unclear mechanisms for symptom alleviation, such as in the case of lithium for bipolar disorder[13] and SSRIs for anxiety and depression[14]. In fact, many psychiatric medications were initially intended for other purposes, and their entrance into psychiatry was only due to the observation of psychiatrically therapeutic side effects in patients and animal models[15]. Owing to an incomplete understanding of the mechanisms underlying these medications or the psychiatric symptoms they treat, many patients are resistant to treatment. For example, reported efficacy rates for antidepressants are around 67%[16]. Efficacy rates for

antipsychotic medications are even lower, near 50%[17]. These suboptimal treatment efficacies are likely due to a mismatch between patient etiology and drug mechanism of action. That is, many patients are likely arriving at their symptoms via different biological routes than the ones that are effectively targeted using current drugs.

This premise forms the basis for the goal of targeted or stratified medicine in psychiatry. This movement aims to parse apart different pathways toward disease[18], with the hope of illuminating new drug targets to better address symptoms in currently underserved patients. Toward this end, there is a strong need to map the full pathways linking common genomic psychiatric risk variants and the symptoms they eventually affect. This commentary aims to briefly outline major questions and current techniques that intend to address this gap in knowledge.

Major questions facing psychiatric genomics

In only two decades since the first published genome-wide association study (GWAS)[19], genomics researchers have amassed an enormous collection of common variants that are associated with altered risk for a host of biological traits. Many researchers are now digging into these lists of loci and attempting to trace the biological pathways taken from risk variants toward disease.

This is a decidedly difficult task. Accurate identification of causal risk loci is substantially complicated by strong linkage disequilibrium (LD), or correlation of alleles in nearby variants due to non-random association, within large blocks across the genome. This renders nearby variants difficult to separate on true causal relationship to risk. To further complicate matters, over 90% of GWAS-identified disease variants lie in noncoding regions of the genome[20], making it difficult to predict their subsequent impact on proteins and downstream biological pathways. GWASs also identify commonly inherited variants that, in disorders that occur before or during the age of reproduction, have not evolved out of the population. This is in contrast to rare, highly penetrant

variants, which give strong risk for disease but which are not maintained in the population at high frequencies. As a result, effect sizes tend to be small for any individual GWAS variant[21]. This has led researchers to hypothesize that GWAS variants tend to contribute disease risk in aggregate or against an epistatic background of genomic risk[22–24]. Taken together, these factors convolute the identification of likely true causal variants out of the massive list of associations that have been generated by GWASs.

Thus, many of the major barriers to a complete understanding of the biological impacts of GWAS variants can be summarized as the following questions: (1) Which human variants are true risk loci, (2) what are the direct biological impacts of these variants, and (3) what are the aggregated biological impacts of these variants. We summarize recent technologies that have attempted to address these questions in an effort to better trace the pathways from apparently low-impact genomic variants all the way to symptoms and severe psychiatric disorders.

Identifying disease-associated loci

To understand the challenges with identifying true causally associated disease loci using GWAS, it is important to understand that GWASs do not query every common variant in the human genome. Due to LD structure across the genome, such a task would be inefficient and highly computationally expensive. Rather, to maintain scalability and feasibility, GWASs take advantage of this known linkage structure to prioritize a limited number of variants that represent overlapping LD blocks from across the genome. Thus, any one variant can serve as a proxy signal for the variants nearby it, allowing researchers to identify loci of interest without explicitly mapping every variant.

This is highly useful for scaling a GWAS to the sizes required for variant detection and simultaneously narrowing the genome search space to prioritized ‘neighborhoods’ of risk. However, a consequence is that GWASs are not easily able to identify which variant or variants

in a linkage block are the causal source of identified risk. Even if tested, variants in high LD, by definition, do not often dissociate. As a result, it is extremely difficult to test for differential effects of two nearby SNPs. Methods have been developed to try to account for this uncertainty and nominate variants to prioritize as hypothesized true sources of disease risk. Chief among these is fine-mapping, which aims to statistically determine the likelihood of causal disease risk for variants in a risk-associated LD block. These fine-mapping efforts then attempt to nominate the variant or variants that are most likely to be the true source of risk in the region. Various methods exist (for review, see Schaid et al. 2018[25]) and are commonly applied in large-scale GWASs to prioritize variants for further study. In the most recent GWAS of schizophrenia from the Psychiatric Genomics Consortium, nearly 300 prioritized loci were fine-mapped for likely causal signals[26] and have been the subject of subsequent research[27,28]. As such, fine-mapping stands as a promising tool to prioritize variants which may be the most likely causal sources of risk in GWAS loci.

GWASs and fine-mapping are highly constrained by genomic LD structure. However, millenia of genetic drift between regionally distinct human populations have resulted in substantial differences in LD structure between people of different ancestries[29], impacting the ability to transport GWAS loci across populations. Further complicating the transracial portability of GWAS results, many of the SNPs used in GWASs have strikingly different allele frequencies across people of different ancestries[30]. This renders SNPs that appear to carry a statistical association with disease risk in one population apparently neutral in another population. As more than 90% of GWASs have been carried out in subjects of European ancestry[31], these differences in population structure impair the utility of European-derived GWAS data for non-European populations[32] and make the identification of true causal variants murky. Encouragingly, many researchers are taking steps to correct this imbalance, conducting GWASs in non-European populations and developing methods to better transport previously

defined GWAS data across different ancestral groups[33]. Further, studying GWASs that were conducted in different ancestries can be extremely useful to narrow the search space for causal variants. Fine-mapping strategies have been developed that take advantage of the different LD blocks present across human populations to help resolve GWAS signals[34,35], assuming the actual true causal variant is still the same across populations. These results highlight the importance of using multi-ancestry GWAS data to identify the true biological signals underlying disease risk and ultimately improve understanding and treatment of disease mechanisms for affected patients of any and all ancestries.

Identifying direct biological impacts of GWAS variants

With the vast majority of GWAS-identified risk variants being located within noncoding regions of the genome, the dominant hypothesis is that these variants alter disease risk by changing gene regulation. Among regulators of gene expression, altered efficacy of enhancers and silencers is often tapped as a potential source of disease risk. *Cis*-regulatory elements (CREs) such as enhancers and silencers regulate the contextually appropriate activation of targeted genes[36,37]. One systematic review of non-coding GWAS variants found that approximately 70% showed evidence of being located within CREs[38]. Candidate CREs are typically identified from epigenomic data. Researchers often look for histone modifications such as H3K27ac[39], frequently in conjunction with data from open chromatin assays like ATAC-seq[40], to map enhancers in the genome. To identify silencers, researchers typically look for histone modifications associated with closed chromatin, like H3K27me3[41], or regions bound by repressive transcription factors such as REST[42].

Once candidate CREs are identified, researchers can test their functional activity in a model system using a CRE-reporter assay. In these assays, genetic constructs are synthesized containing a reporter gene under the control of a minimal promoter, which is further regulated by

a hypothesized CRE. These assays are highly scalable and enable testing of up to several thousand candidate CREs in a single experiment, earning them the moniker *massively parallel* reporter assays (MPRAs)[43]. Since their advent, MPRAS have also been adapted to test post-transcriptional regulatory elements, including 5' UTR variants[44], 3' UTR variants[45], and splice site variants[46]. As a powerful and versatile technique, MPRAs can be used to directly test the hypothesis that GWAS-identified risk variants alter transcriptional regulatory function in the model system of interest.

Recent applications of MPRAs have tested whether GWAS-identified psychiatric risk variants alter CRE activity in model organisms. One recent investigation tested over 5,000 fine-mapped schizophrenia variants in human neural progenitor cells and identified 439 variants with evidence for allelic differences in enhancer activity[47]. Others have used MPRAs to identify silencers which appear to be disrupted by psychiatric and neurological risk variants[48]. Animal models have also been used for *in vivo* testing of CRE activity, taking advantage of the ability to model activity during biological contexts that are not possible to recreate *in vitro*. For example, one large screen tested over 1,000 variants from major depressive disorder (MDD) risk loci in postnatal mouse brain. Results found a strong impact of sex on allelic differences in enhancer activity[49]. Subsequent analyses showed that many alleles likely disrupted the transcription factor binding motifs of sex hormone receptors. This nominated a plausible mechanism by which the observed sex difference in CRE activity might have occurred, and which may be related to the sex disparity[50] in MDD prevalence in humans. Results from this study implied that GWAS risk variants within CREs likely altered the appropriate transcriptional regulation *in vivo*, highlighting a mechanism by which these variants may alter biology and increase risk for disease.

To identify exactly which genes may be aberrantly regulated by CREs containing risk-associated variants, many researchers look to expression quantitative trait loci (eQTL) or chromatin accessibility quantitative trait loci (caQTL) data. eQTLs are variants that are

correlated with altered gene expression levels[51], while caQTLs are variants that are correlated with altered chromatin accessibility[52]. Both measures are used to identify variants that may alter the function of transcriptional regulators, though they are subject to the same limitations of LD and variant correlation structure that exists in GWASs. Many GWAS variants have been identified as eQTLs or caQTLs in postmortem human brain tissue[52,53]. This further supports the hypothesis that risk alleles may disrupt appropriate *cis*-regulatory activity. By identifying the gene targets whose expression is thought to be regulated by these risk-associated CREs (eGenes), researchers can model the transcriptional impact of CRE-disrupting variants. Recent work to this end used CRISPR-dCas9 to mimic eGene expression from GWAS variants in 2D human neuronal models. Results highlighted synaptic and transcriptional impacts as convergent pathways across a range of eGene perturbations[54]. Findings such as these can be combined with MPRAs, epigenomic assays, and eQTL/caQTL data to assign likely regulatory and transcriptomic effects onto identified GWAS variants. Subsequently, researchers can study the putative molecular impacts of these previously cryptic noncoding risk loci.

Identifying aggregated biological impacts of GWAS variants

One challenge of interpreting the biological effects of GWAS risk variants is that GWAS-identified variants typically have low effect sizes for disease risk. This is because GWASs are not typically intended for detection of rare, highly penetrant risk variants (minor allele frequency [MAF] < 1%), in large part due the need for unsustainably large sample sizes to detect rare variants for disease via GWAS (for discussion, see Abdellaoui et al. 2023[55]). However, genomic variants of high effect sizes for severe diseases that occur during or before reproductive ages are typically these rarest variants, as they are subject to negative selection[21,56]. As a result, GWASs typically query variants of individually low effect size, as these variants tend to be more common in the general population (MAF ≥ 1%) for diseases that may impact sexual selection and reproduction. Consequently, researchers widely hypothesize a cumulative model of disease

risk for many GWAS-identified variants. In these models, variants work against an epistatic background of genomic risk, summing or working synergistically to increase risk for diagnostically relevant symptoms[22–24].

To test the effects of cumulative genomic risk, researchers can group variants together to ask questions about the aggregated impact of having many risk variants at once. A highly useful tool for this is the polygenic risk score (PGS), which measures an individual's cumulative genomic risk for a trait of interest based on their GWAS-identified variants[24]. Traditional PGSs aggregate an individual's risk from across their genome and can predict increased risk for diseases, symptoms, and biomarkers. For example, recent PGS studies in schizophrenia have not only identified individuals at heightened risk for schizophrenia[57], but have also identified associations between genome-wide schizophrenia risk and treatment response[58], as well as neurological and cognitive measures[59,60]. These results highlight the power of considering low-effect variants in aggregate in psychiatric disease.

An updated take on the PGS is the pathway PGS (pPGS). pPGSs first subset GWAS variants into those thought to affect genes within biological pathways of interest. Subsequently, polygenic risk is calculated individually within each pathway. This allows researchers to directly query the clinical outcome of accumulated genomic impacts onto an etiological pathway of interest. A newer approach, results from studies utilizing pPGS are already showing promise in helping trace pathways from GWAS risk toward disease symptoms and outcomes. One exciting implementation of pPGS research in psychotic disorders identified subjects with schizophrenia and bipolar disorder who have genomic risk concentrated in pathways that are potentially targetable via drug repurposing[61]. Another recent publication found associations between specific endophenotypes of psychotic disorders and genomic risk hypothesized to impact different synapse types[57]. Pathway pPGSs are a powerful tool in psychiatric research, enabling

researchers to directly query the clinical impacts of altered biology associated with multiple convergent GWAS risk variants.

Bringing these together

A stratified medicine approach has already shown success in diseases with better understood biology than most psychiatric disorders. For example, a recent machine-learning effort in a large cohort of subjects with diabetes used biomarkers and clinical data to identify five subclusters of patients[62]. These subgroups replicated in independent cohorts and showed distinct profiles of genetic risk. Interestingly, the groups also tended to have undergone distinct paths of pharmacological treatment. Results like this highlight the potential clinical utility of pathway-based subgrouping and show promise for offering distinct treatments to patients with distinct mechanisms of disease.

This biologically clustered type of treatment is the goal for many researchers and clinicians in psychiatry. Once common risk variants for disease have been correctly identified, annotated with a biochemical function, and partitioned into their respective etiological pathways, researchers and clinicians will have a more straightforward path toward tracing the mechanisms that link genomic variants to disease. Improved mapping of common risk variants from GWAS onto their affected biological pathways will facilitate better partitioning of psychiatric symptoms into their respective etiologies. Eventually, this will allow for the development and application of treatments in a stratified manner to intentionally target disrupted biological pathways, enabling treatments, preventions, and potentially even cures for patients suffering from psychiatric disease.

References

1. Scadding, J. G. The semantic problems of psychiatry. *Psychol. Med.* **20**, 243–248 (1990).

2. Cooper, J. Disorders are different from diseases. *World Psychiatry* **3**, 24 (2004).
3. Kendler, K. S., Tabb, K. & Wright, J. The Emergence of Psychiatry: 1650–1850. *AJP* **179**, 329–335 (2022).
4. Roberts, W. *et al.* Across the centuries: Piecing together the anatomy of the heart. *Translational Research in Anatomy* **17**, 100051 (2019).
5. Handley, S. L. & Mithani, S. Effects of alpha-adrenoceptor agonists and antagonists in a maze-exploration model of ?fear?-motivated behaviour. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **327**, 1–5 (1984).
6. Nadler, J. J. *et al.* Automated apparatus for quantitation of social approach behaviors in mice. *Genes Brain and Behavior* **3**, 303–314 (2004).
7. Weeks, J. R. Experimental Morphine Addiction: Method for Automatic Intravenous Injections in Unrestrained Rats. *Science* **138**, 143–144 (1962).
8. Ennaceur, A. & Meliani, K. A new one-trial test for neurobiological studies of memory in rats. III. Spatial vs. non-spatial working memory. *Behavioural Brain Research* **51**, 83–92 (1992).
9. Chadman, K. K., Yang, M. & Crawley, J. N. Criteria for validating mouse models of psychiatric diseases. *Am J Med Genet B Neuropsychiatr Genet* **150B**, 1–11 (2009).
10. Hinzen, W. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. *Front. Psychol.* (2015) doi:10.3389/fpsyg.2015.00971.
11. Bandettini, P. A. Twenty years of functional MRI: The science and the stories. *NeuroImage* **62**, 575–588 (2012).
12. Newson, J. J. & Thiagarajan, T. C. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Front. Hum. Neurosci.* **12**, 521 (2019).
13. Anand, A. *et al.* Integrative analysis of lithium treatment associated effects on brain structure and peripheral gene expression reveals novel molecular insights into mechanism of action. *Transl Psychiatry* **10**, 103 (2020).
14. Taylor, C., Fricker, A. D., Devi, L. A. & Gomes, I. Mechanisms of action of antidepressants: from neurotransmitter systems to signaling pathways. *Cellular Signalling* **17**, 549–557 (2005).
15. Ban, T. A. The role of serendipity in drug discovery. *Dialogues in Clinical Neuroscience* **8**, 335–344 (2006).
16. Santarsieri, D. & Schwartz, T. Antidepressant efficacy and side-effect burden: a quick guide for clinicians. *DIC* **4**, 1–12 (2015).
17. Leucht, S. *et al.* Sixty Years of Placebo-Controlled Antipsychotic Drug Trials in Acute Schizophrenia: Systematic Review, Bayesian Meta-Analysis, and Meta-Regression of Efficacy Predictors. *AJP* **174**, 927–942 (2017).
18. Schumann, G. *et al.* Stratified medicine for mental disorders. *European Neuropsychopharmacology* **24**, 5–50 (2014).
19. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat Commun* **11**, 5900 (2020).
20. Peña-Martínez, E. G. & Rodríguez-Martínez, J. A. Decoding Non-coding Variants: Recent Approaches to Studying Their Role in Gene Regulation and Human Diseases. *Front. Biosci. (Schol Ed)* **16**, 4 (2024).
21. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25

UK Biobank traits reveals action of negative selection. *Nat Commun* **10**, 790 (2019).
22. Andreasen, N. C. *et al.* Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes. *Mol Psychiatry* **17**, 1093–1102 (2012).
23. Webber, C. Epistasis in Neuropsychiatric Disorders. *Trends in Genetics* **33**, 256–265 (2017).
24. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 44 (2020).
25. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491–504 (2018).
26. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
27. Kraft, J. *et al.* Identifying drug targets for schizophrenia through gene prioritization. Preprint at https://doi.org/10.1101/2024.05.15.24307423 (2024).
28. Dong, P., Voloudakis, G., Fullard, J. F., Hoffman, G. E. & Roussos, P. Convergence of the dysregulated regulome in schizophrenia with polygenic risk and evolutionarily constrained enhancers. *Mol Psychiatry* **29**, 782–792 (2024).
29. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477–485 (2008).
30. Mattei, J. *et al.* Disparities in allele frequencies and population differentiation for 101 disease-associated single nucleotide polymorphisms between Puerto Ricans and non-Hispanic whites. *BMC Genet* **10**, 45 (2009).
31. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242–243 (2020). (Accessed July 6, 2024)
32. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun* **11**, 3865 (2020).
33. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).
34. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *The American Journal of Human Genetics* **97**, 260–271 (2015).
35. Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet* **51**, 1670–1678 (2019).
36. Nord, A. S. & West, A. E. Neurobiological functions of transcriptional enhancers. *Nature Neuroscience* (2020) doi:10.1038/s41593-019-0538-5.
37. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends in Genetics* **37**, 514–527 (2021).
38. Alsheikh, A. J. *et al.* The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. *BMC Med Genomics* **15**, 74 (2022).
39. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155**, 1521–1531 (2013).
40. Stępniak, K. *et al.* Mapping chromatin accessibility and active regulatory elements

reveals pathological mechanisms in human gliomas. *Nat Commun* **12**, 3621 (2021).
41. Zhang, Y.-Z. *et al.* Coupling of H3K27me3 recognition with transcriptional repression through the BAH-PHD-CPL2 complex in Arabidopsis. *Nat Commun* **11**, 6212 (2020).
42. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**, 1061 (2020).
43. McAfee, J. C. *et al.* Focus on your locus with a massively parallel reporter assay. *J Neurodevelop Disord* **14**, 50 (2022).
44. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37**, 803–809 (2019).
45. Griesemer, D. *et al.* Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247-5260.e19 (2021).
46. Chong, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Molecular Cell* **73**, 183-194.e8 (2019).
47. McAfee, J. C. *et al.* Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. *Cell Genomics* **3**, 100404 (2023).
48. Huang, D. & Ovcharenko, I. The contribution of silencer variants to human diseases. Preprint at https://doi.org/10.1101/2024.05.17.24307558 (2024).
49. Mulvey, B., Selmanovic, D. & Dougherty, J. D. Sex Significantly Impacts the Function of Major Depression–Linked Variants In Vivo. *Biological Psychiatry* **94**, 466–478 (2023).
50. Li, S. *et al.* Sex difference in incidence of major depressive disorder: an analysis from the Global Burden of Disease Study 2019. *Ann Gen Psychiatry* **22**, 53 (2023).
51. Zhang, J. & Zhao, H. eQTL studies: from bulk tissues to single cells. *Journal of Genetics and Genomics* **50**, 925–933 (2023).
52. Zeng, B. *et al.* Genetic regulation of cell type–specific chromatin accessibility shapes brain disease etiology. *Science* **384**, eadh4265 (2024).
53. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442–1453 (2016).
54. Townsley, K. G. *et al. Convergent Impact of Schizophrenia Risk Genes*. http://biorxiv.org/lookup/doi/10.1101/2022.03.29.486286 (2022) doi:10.1101/2022.03.29.486286.
55. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 179–194 (2023).
56. Fiziev, P. P. *et al.* Rare penetrant mutations confer severe risk of common diseases. *Science* **380**, eabo1131 (2023).
57. Warren, T. L. *et al.* Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis. *Molecular Psychiatry* (2024) doi:10.1038/s41380-024-02457-0.
58. Zhang, J.-P. *et al.* Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *AJP* **176**, 21–28 (2019).
59. Shafee, R. *et al.* Polygenic risk for schizophrenia and measured domains of cognition in individuals with psychosis and controls. *Transl Psychiatry* **8**, 78 (2018).
60. He, Q. *et al.* Influence of polygenic risk scores for schizophrenia and resilience on the

cognition of individuals at-risk for psychosis. *Transl Psychiatry* **11**, 518 (2021).
61. Reay, W. R. *et al.* Genetics-informed precision treatment formulation in schizophrenia and bipolar disorder. *The American Journal of Human Genetics* **109**, 1620–1637 (2022).
62. Ahlqvist, E. *et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology* **6**, 361–369 (2018).

c. Contributions of this dissertation toward mapping GWAS variants onto psychiatric etiologies

In this dissertation, I present three chapters which aim to support the identification of biological pathways impacted by GWAS-identified psychiatric risk variants.

In Chapter 1, I describe in-depth a powerful method which is used to directly test hypothesized biological function of CREs in a model organism: the *in vivo* CRE-reporter assay. Common risk variants for psychiatric disorders are highly enriched in regions which show epigenetic evidence of CRE activity[1]. This has led to the hypothesis that risk variants may disrupt the ability of CREs to regulate gene expression in a contextually appropriate manner. To directly test the hypothesis that these risk regions act as CREs, these genomic elements can be tested for their ability to induce or suppress expression of a reporter gene. In Chapter 1, I detail methodologies to perform this assay to test enhancers in postnatal mouse brain. I further provide an example of these experiments in which I tested enhancer activity of a region containing a schizophrenia risk variant in one of the top schizophrenia risk genes, *CACNA1C*. I additionally validated activity of this and two other enhancers *in vitro* in HEK-293T cells. My efforts resulted in one first-author publication in the *Journal of Visualized Experiments* (2022)[2], and my contributions to one main figure and one supplementary figure in an additional publication from our lab in *eLife* (2021)[3] on which I am fourth author.

In Chapter 2, I extended my deployment of the massively parallel reporter assay (MPRA) to test for allelic changes in *cis*-regulatory activity at a prioritized set of schizophrenia-associated

variants in postnatal mouse brain. I first describe major redesigns I made to our lab's MPRA protocol to minimize confounds and improve scientific rigor. These efforts substantially enhanced the sensitivity of our assay to detect allelic differences, enabling not only the identification of CREs, but also the direct testing of functional impacts of the risk SNP. I then describe the execution and results of my MPRA, assessing allelic differences at highlighted CREs as well as longitudinal changes between juvenile and adolescent mice. I contextualize my findings with epigenetic and gene expression data that have been published by other researchers, generating hypotheses on potential gene targets of CREs that were identified as active in my assay. I recommend direct testing of these hypotheses in future research, which may help map the SNPs which showed impacts on CRE activity onto their subsequently affected biological pathways.

In Chapter 3, I aggregated GWAS-identified common risk variants into hypothesized pathways of risk by dividing them into variants predicted to impact genes involved in four major synapse types in the brain: glutamate, GABA, dopamine, and serotonin synapses. I describe research which I led in a multi-lab collaboration where we used pathway polygenic risk scores to identify psychosis subjects with heightened risk in these pathways. This work identified endophenotypes of psychotic disorders which were associated with distinct profiles of polygenic risk and which showed different clinical outcomes. My work on this project resulted in a co-first author publication in *Molecular Psychiatry* (2024)[4].

Together, these projects aim to (Ch. 1 & 2) append direct biological function onto non-coding psychiatric risk variants, and (Ch. 3) identify stratifiable subtypes of psychotic disorders which are associated with distinct biological pathways onto which genomic risk variants may converge. My work has improved the understanding of the relationship between genomic risk and psychiatric etiologies. Hopefully, these efforts will contribute to a biologically stratified approach

to therapeutics in psychiatric disorders and improve outcomes for patients with psychiatric symptoms.

References

1. Zeng, B. *et al. Genetic Regulation of Cell-Type Specific Chromatin Accessibility Shapes the Etiology of Brain Diseases*. http://biorxiv.org/lookup/doi/10.1101/2023.03.02.530826 (2023) doi:10.1101/2023.03.02.530826.
2. Warren, T. L., Lambert, J. T.* & Nord, A. S.* AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *JoVE* 62650 (2022) doi:10.3791/62650.
* indicates co-senior authors
3. Lambert, J. T.* & Su-Feher, L.* *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
* indicates co-first authors
4. Warren, T. L.* & Tubbs, J. D.* *et al.* Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis. *Molecular Psychiatry* (2024) doi:10.1038/s41380-024-02457-0.
* indicates co-first authors

Chapter 1:

Development and deployment of *in vivo* MPRA to test enhancer activity in postnatal mouse brain

a. Preface

In the present chapter, I describe the development and deployment of an *in vivo* enhancer-reporter expression assay using an adeno-associated virus (AAV) in postnatal mouse brain. Enhancer-reporter testing is a powerful assay for directly testing the functional activity of putative enhancer elements.

Enhancers are transcription factor-binding sites in DNA that are among the chief regulators of gene expression in a cell[1]. Given their importance for gene expression, disruption to enhancers has been strongly linked to disease. As such, functional enhancer testing is a popular approach for researchers who study common risk variants[1–3]. Researchers frequently deploy assays testing enhancer elements *in vitro*[4,5], given the fast and straightforward nature of working with immortalized cell lines such as HEK cells. While this technique can successfully identify enhancer elements in cases where the appropriate transcription factors are present in the cell, *in vitro* assays can miss some of the most important nuances of enhancer function in the tissues and cells in which they are natively active – enhancers not only regulate gene expression, but they do so in a manner that is highly specific to the precise type and developmental stage of the cell, tissue, and organism they are in[1]. As *in vitro* systems cannot accurately model the enormous complexity of biology, development, and experience to which a living brain is exposed, there is a need for technologies that can test enhancer activity *in vivo* in a model organism.

In the publication comprising the first portion of this chapter, I describe the methods development and deployment of such an assay in our lab. I showed robust activity of hypothesized enhancer elements at multiple stages of development in postnatal mouse brain. The method described herein forms the basis of multiple ongoing projects in our lab, including work of mine that is described in the second chapter of my dissertation.

I was first author on this methods manuscript, which was published in the *Journal of Visualized Experiments* on March 31, 2022[6]. I helped refine and optimize this technology in our lab, executed the representative experiments with the exception of figure panel 4A, and was the primary writer for the manuscript text. The resulting publication and all work developing and refining *in vivo* AAV-based enhancer-reporter assays in our lab was conducted in strong collaboration with our lab's former postdoc Jason Lambert, who is credited as co-senior author on the manuscript.

I subsequently include an appendix in which I describe my use of this methodology to validate activity at a top enhancer candidate identified by our lab in an *in vivo* MPRA in P7 mouse brain. The MPRA was published in *eLife* on November 9, 2021[7]. I contributed experiments which are featured in Figure 4 and its second corresponding supplementary figure, and I am fourth-author on the paper. My work used an AAV-based enhancer-reporter assay to validate activity through P28 at an enhancer which contains a schizophrenia risk SNP in the top schizophrenia risk gene *CACNA1C*. I additionally include unpublished results from *in vitro* validation experiments in HEK293T cells at this enhancer and two other nearby enhancers in *CACNA1C* which contain schizophrenia risk SNPs.

These *in vivo* methods were still in development when I joined our lab and their optimization and validation constituted much of my work during my first two years. In Chapter 2, I describe how I

further expanded and improved upon these methods to launch a larger, longitudinal MPRA of schizophrenia risk SNPs in postnatal mouse brain.

References

1. Nord, A. S. & West, A. E. Neurobiological functions of transcriptional enhancers. *Nature Neuroscience* (2020) doi:10.1038/s41593-019-0538-5.
2. Alsheikh, A. J. *et al.* The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. *BMC Med Genomics* **15**, 74 (2022).
3. Zeng, B. *et al. Genetic Regulation of Cell-Type Specific Chromatin Accessibility Shapes the Etiology of Brain Diseases*. http://biorxiv.org/lookup/doi/10.1101/2023.03.02.530826 (2023) doi:10.1101/2023.03.02.530826.
4. Xiao, X., Chang, H. & Li, M. Molecular mechanisms underlying noncoding risk variations in psychiatric genetic studies. *Mol Psychiatry* **22**, 497–511 (2017).
5. Farrow, S. L. *et al.* Identification of 27 allele-specific regulatory variants in Parkinson's disease using a massively parallel reporter assay. *npj Parkinsons Dis.* **10**, 44 (2024).
6. Warren, T. L., Lambert, J. T.* & Nord, A. S.* AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *JoVE* 62650 (2022) doi:10.3791/62650.
* indicates co-senior authors
7. Lambert, J. T.* & Su-Feher L.* *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
* indicates co-first authors
8. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**, 402–408 (2001).
9. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* **5**, 99–114 (1949).

b. Resulting Publication

DOI: https://dx.doi.org/10.3791/62650

J. Vis. Exp. (181), e62650, doi:10.3791/62650 (2022).

Title:

AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain

Author list:

Tracy L. Warren[1,2], Jason T. Lambert[1,2*], Alex S. Nord[1,2*]

* Indicates co-senior authors

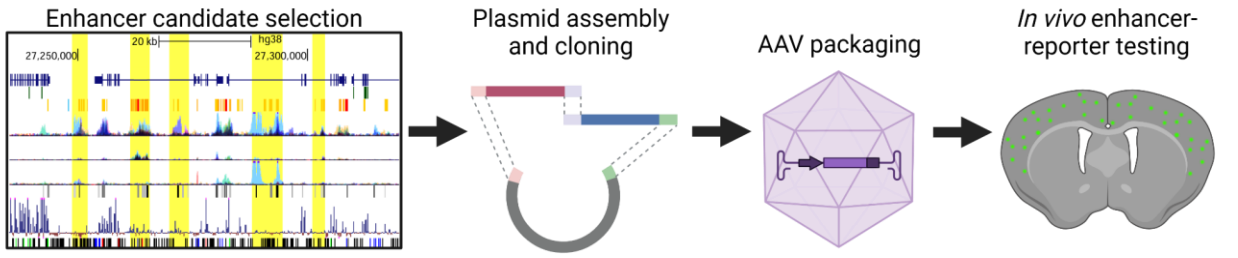
Author affiliations:

[1]Department of Psychiatry and Behavioral Sciences, University of California, Davis,

[2]Department of Neurobiology, Physiology and Behavior, University of California, Davis

Correspondence to Alex S. Nord (asnord@ucdavis.edu)

Graphical abstract



Created with BioRender.com

Abstract

Enhancers are binding platforms for a diverse array of transcription factors that drive specific expression patterns of tissue- and cell-type-specific genes. Multiple means of assessing non-coding DNA and various chromatin states have proven useful in predicting the presence of enhancer sequences in the genome, but validating the activity of these sequences and finding the organs and developmental stages they are active in is a labor-intensive process. Recent advances in adeno-associated virus (AAV) vectors have enabled the widespread delivery of transgenes to mouse tissues, enabling *in vivo* enhancer testing without necessitating a transgenic animal. This protocol shows how a reporter construct that expresses EGFP under the control of a minimal promoter, which does not drive significant expression on its own, can be used to study the activity patterns of candidate enhancer sequences in the mouse brain. An AAV-packaged reporter construct is delivered to the mouse brain and incubated for 1–4 weeks, after which the animal is sacrificed, and brain sections are observed under a microscope. EGFP appears in cells in which the tested enhancer is sufficient to initiate gene expression, pinpointing the location and developmental stage in which the enhancer is active in the brain. Standard cloning methods, low-cost AAV packaging, and expanding AAV serotypes and methods for *in*

vivo delivery and standard imaging readout make this an accessible approach for the study of how gene expression is regulated in the brain.

Introduction

Enhancers are genomic *cis*-regulatory elements that serve as transcription factor binding sites and can drive the expression of a target gene in a spatiotemporally specific manner[1,2]. They are differentially active in different cell types, tissues, and stages of development and can be substrates of disease risk-related genomic variation[3, 4]. Thus, the need to understand the dynamics of enhancer function is critical to progress in both translational and basic science applications within genomics. In silico predictions of enhancer activity can serve as excellent resources for generating hypotheses as to enhancer capability[5, 6]. Such predicted enhancer activity can require additional validation and interrogation for a full understanding of the functional activity. Enhancer reporter assays have proved valuable for this purpose across a variety of systems, from cells to animals[7, 8, 9]. Towards extending these studies in a flexible and cost-effective transient *in vivo* context, this protocol describes the use of *in vivo* AAV-based methods to test putative enhancer sequences for their ability to drive the expression of an ectopic reporter gene in the postnatal mouse brain. This family of methods has utility for interrogating single candidate sequences or parallel library screening and is relevant for basic and translational research.

This method combines in a single plasmid a putative enhancer candidate DNA sequence with a reporter gene (here EGFP), under the control of a minimal promoter that alone does not drive significant expression. The plasmid is packaged into recombinant AAV (rAAV) and injected into an animal model. While the application here is to the brain, various rAAV serotypes enable infection across different tissue types so that this approach can be extended to other systems[10]. After a period of time, the brain can be collected and assayed for the expression of the reporter

gene. Strong expression, compared with controls, indicates that the tested candidate sequence was able to “enhance” the expression of the gene (Figure 1). This simple design offers an easy and clear approach to test a sequence for enhancer activity *in vivo* in the brain.

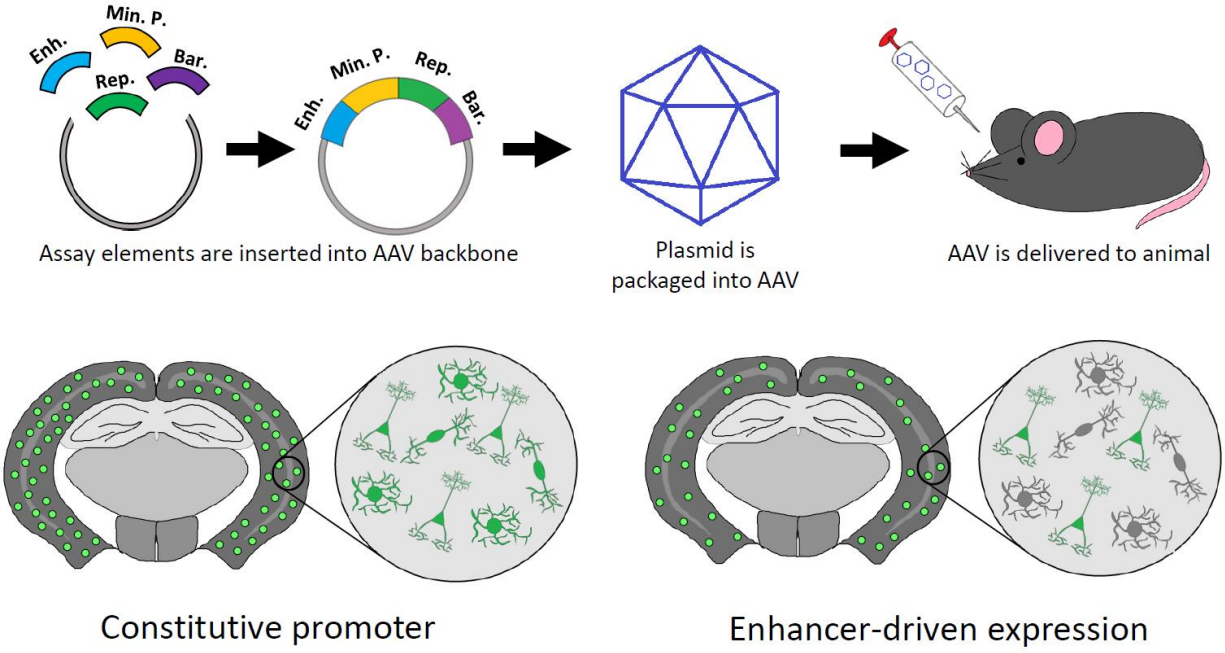


Figure 1: Overview of the protocol.

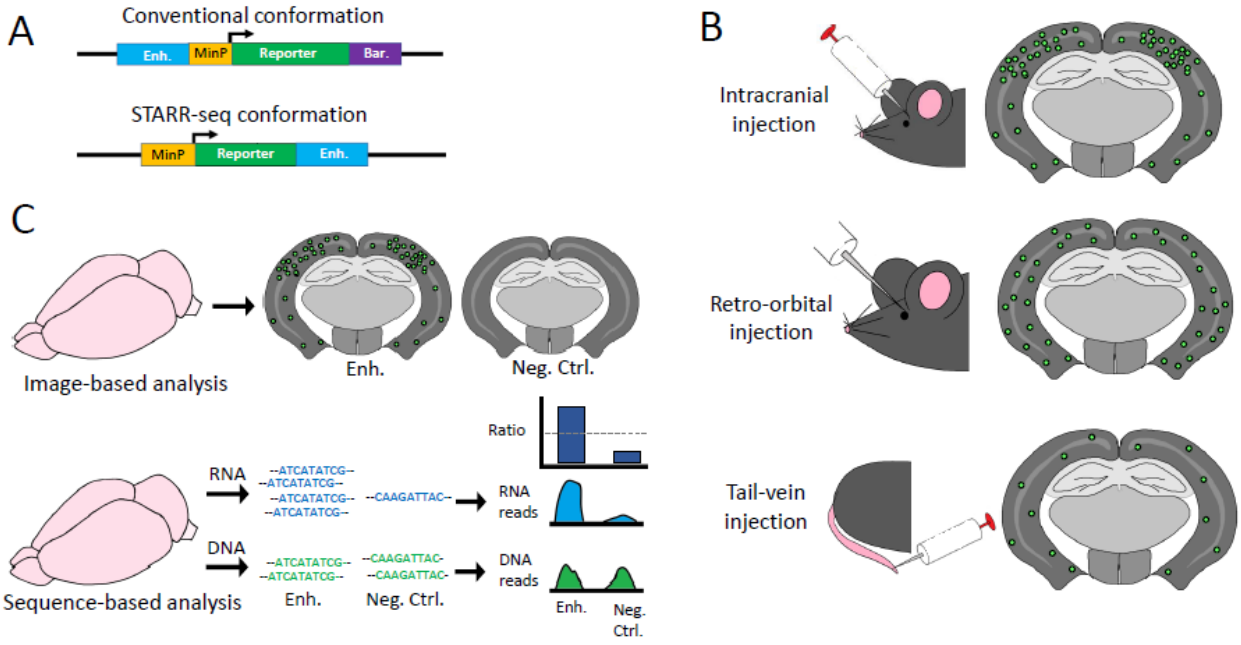
Key elements of this assay include a candidate enhancer element, a minimal promoter, a reporter gene, and, depending on assay design, a barcode sequence for unique tagging. These elements are cloned into an AAV plasmid backbone and packaged into AAV. The virus is delivered to the animal and left to incubate for a number of days or weeks. Finally, the tissue of interest is collected and enhancer activity is ascertained via imaging or sequencing. Enhancer-driven transgene expression can produce expression with cell-type, regional, and temporal specificity, in contrast to constitutive drivers.

In addition to testing for enhancer capability of a sequence, this method can be combined with techniques to determine cell-type enhancer activity. In sequence-based approaches to determining differential enhancer activity, sorting cells on cell-type-specific markers prior to DNA

and RNA sequencing can allow researchers to determine if different cell types show differential enhancer activity, as was described in Gisselbrecht et al.[11]. In imaging-based approaches, co-labeling images with cell-type-specific markers allows examination of whether cells exhibiting enhancer-driven fluorescence also display cell-type markers of interest[12, 13, 14, 15, 16]. Enhancer reporter assays enable direct testing of risk-associated allelic variation in enhancers for effects on enhancer capability. The vast majority of risk loci identified in genome-wide association studies (GWAS) lie in non-coding regions of the genome[17]. Functional annotation studies of these risk loci indicate that a large portion likely act as enhancers[18, 19, 20]. MPRA deployment *in vivo* can allow testing of these risk-associated variants for enhancer activity in brain[12, 21]. Finally, delivery and collection at different time points can offer insights into the developmental stages during which an enhancer is active.

Enhancer-reporter plasmid designs are diverse and can be customized to suit experimental goals. There are several options for minimal promoters that have been used in enhancer research, such as the human β-globin minimal promoter[22] and the mouse Hsp68 minimal promoter[23]. These promoters are known to drive low levels of expression unless coupled with an enhancer element to activate them. In contrast, constitutive promoter elements drive strong expression of the transgene, useful for positive control or to test for enhancer function against a background of robust expression. Common choices for constitutive promoters include CAG, a hybrid promoter derived from the chicken β-actin promoter and the cytomegalovirus immediate-early enhancer[24], or human EF1α[25]. Since enhancers are known to work bidirectionally[26], the orientation and location of the enhancer relative to the minimal promoter are flexible (Figure 2A). Traditional enhancer-reporter assays place the enhancer upstream of the promoter and, in library deliveries, include a barcode sequence downstream of the reporter gene to associate sequencing reads with the tested enhancer[27]. However, enhancers can also be placed in the open reading frame of the reporter gene and serve as their own barcode sequence, as is done

in STARR-seq[28]. The protocol described here utilizes the STARR-seq assay design, placing the candidate enhancer sequence into the 3' UTR of the reporter gene. While the STARR-seq orientation offers the benefit of more streamlined cloning, it is less well understood than the conventional approach and may induce variable RNA stability between constructs. The described methods can be easily adapted to either the STARR-seq or conventional orientation with minor alterations to the cloning process that have been described elsewhere[27, 29].



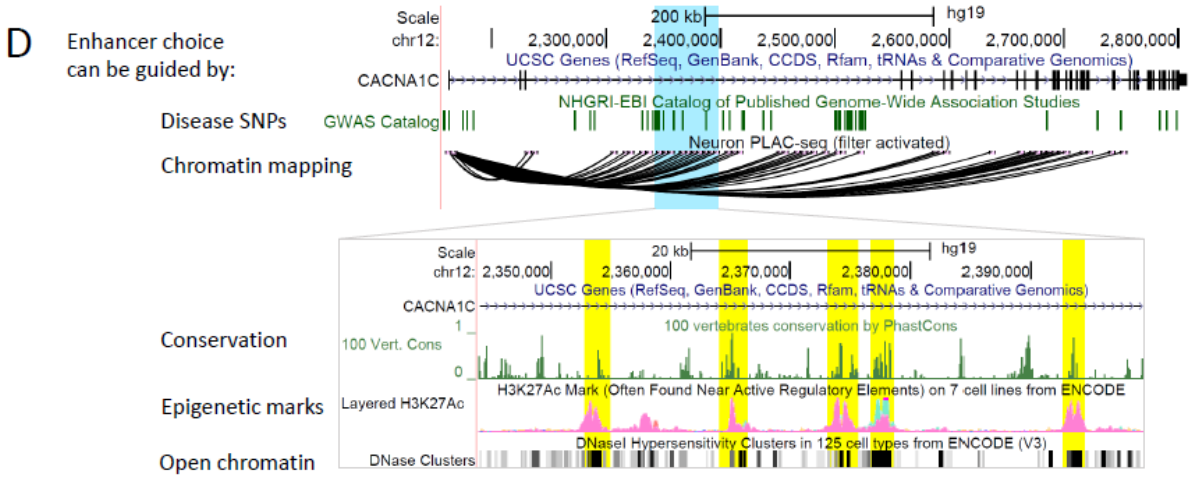


Figure 2: Flexibility in assay design.

(A) Plasmid orientation can place the enhancer upstream of the minimal promoter or downstream of the reporter gene, as in STARR-seq[28]. For sequence-based readouts, a barcode is included downstream of the reporter gene, or the enhancer can serve as its own barcode in the second orientation. (B) Common viral delivery techniques to the brain include intracranial injections, retro-orbital injections, or tail-vein injections, which may result in different densities of transduced cells in different tissues. (C) Readouts can be sequencing-based or imaging-based. In sequencing-based readouts, enhancer activity is defined by a relative increase of RNA sequence versus input DNA sequence (controlling for AAV delivery). In imaging readouts, expression of the reporter gene is increased where the enhancer is active. (D) Potential intronic enhancers in the psychiatric risk-associated gene *CACNA1C* are highlighted. In the top image, a broad region is selected that shows strong associations to GWAS-tested traits and interactions with the *CACNA1C* promoter via PLAC-seq[20]. In the inset, smaller candidate regions are identified that are evolutionarily conserved at the sequence level, are in areas of open chromatin, and show epigenetic marks indicative of enhancers.

Different methods of AAV delivery can be employed to further customize this technique to fit experimental goals (Figure 2B). Direct intracranial injections, described further in this protocol, deliver a high concentration of virus directly to the brain[30]. This gives a high transduction efficiency centered at the site of injection, making this an excellent technique for experiments looking to maximize the density of transduced cells over an area of tissue. Stereotactic injection can help standardize the site of injection across animals for reproducible localized transduction. Intracranial injections are most straightforward in early postnatal animals. As an alternative technique, systemic injections can deliver transgenes using AAVs with serotypes capable of crossing the blood-brain barrier[31]. Tail-vein injections allow the virus to circulate throughout the

body, enabling generalized delivery across many tissues[10]. Retro-orbital injections are another systemic injection technique that delivers the virus behind the eye into the retro-orbital sinus[32]. This offers a more direct route for the AAV from the venous system to the brain, resulting in a higher concentration of transduced cells in the brain than injections into more peripheral blood vessels[33].

Another flexible aspect of this technique is the method of readout. Broadly, options can be described as reporter-based or sequencing-based (Figure 2C). Incorporating a fluorescent reporter such as GFP into the open reading frame of the construct results in the expression of the fluorescent protein in any transduced cells where the candidate enhancer drove expression. Labeling and imaging techniques such as immunohistochemistry enable signal amplification. Sequencing-based readout techniques involve identifying sequences from the delivered construct in RNA collected from the tissue. By quantifying the amount of viral DNA that was initially delivered, the comparison of expressed RNA versus delivered DNA can be used to determine the degree to which a tested enhancer sequence was capable of driving increased expression of the transgene, for example, in the context of a massively parallel reporter assay (MPRA). MPRAs offer a powerful expansion of these techniques to test up to thousands of candidate enhancers for activity simultaneously and have been described extensively in genomics research[12, 27, 34, 35, 36]. Higher throughput screening is achieved by executing cloning, packaging, delivery, and sequencing steps for candidate enhancers in batch rather than individually.

Candidate enhancer selection provides another opportunity for flexibility (Figure 2D). For example, this assay can be used to identify enhancers of a specific gene, to ascertain the function of non-coding DNA regions of interest, or to determine specific cell types or developmental stages during which an enhancer is active - all of which serve goals both in basic science and in disease research. Generally, candidate enhancer selection is driven by in silico

predictions of enhancer activity. Commonly, in silico predictions include ChIP-seq for histone modifications that indicate likely enhancers, such as H3K27ac[37] and chromatin accessibility mapping[38]. Finally, a growing area of research is the function-based screening of synthetically designed enhancer elements, enabling studies of how enhancer sequence directs function[39] and the design of enhancers with specific properties[40].

Protocol

This protocol has been approved by the UC Davis Institutional Animal Care and Use Committee (Protocol #22339) and the UC Davis Institutional Biosafety Committee (BUA-R1903). This protocol has been tested on C57BL/6J mice of both sexes at postnatal day 0–1.

1. Clone the enhancer candidate sequence into the AAV vector plasmid.

NOTE: The representative protocols are given, but the cloning strategy has a high degree of flexibility.

1. Select or design a reporter construct. Here, the enhancer candidate is inserted into the 3' untranslated region (UTR) of the EGFP reporter gene, which is expressed under the control of the Hsp68 minimal promoter.

   NOTE: Many MPRA plasmid constructs appropriate for this assay have been deposited to Addgene[41] and are available for research use.
2. Design PCR primers to amplify a sequence of interest. Add to the 5' end of the primers the appropriate homology arm for Gibson cloning (~35 bp corresponding to sequence 5' upstream of the insertion location, top strand for the forward primer, and bottom strand for the reverse primer).

   NOTE: Ensure that the base primer sequence is ~18–24 bp long and ~50%–60% GC content with a melting temperature of approximately 57–59 °C.
3. Perform PCR from a DNA sample using the designed primers.
   1. Assemble the PCR reaction mix in 0.2 mL PCR tubes as described in Table 1.2.

2. Place the PCR reaction mix(es) on a thermal cycler and cycle according to the program mentioned in Table 1.

Table 1: PCR reaction mix and thermocycling conditions.

PCR Reaction Mix			
5x reaction buffer		10 µL	
dNTPs		final concentration of 200 µM each	
Cloning primers		final concentration of 0.2 µM each	
Template DNA		50 ng	
High fidelity polymerase		final concentration of 0.02 U/µL	
Nuclease-free water		to reach 50 µL final reaction volume	
Thermocycling conditions			
30 cycles of the following:			Notes
Step	Temperature	Time	
Denature	98 °C	10 s	
Anneal	60 °C	20 s	Optimal temperature may vary based on primers.
Extend	72 °C	30 s	Optimal temperature and duration may vary based on polymerase.
Final Extension	72 °C	2 min	Optimal temperature and duration may vary based on polymerase.
Hold	4 °C		

3. Confirm the success of PCR by running 5 µL of PCR product on a 0.7%–1% agarose gel at 100 V for 30–60 min. Check for the fragment of predicted length.
4. Cleanup the PCR product using a commercial enzymatic reaction cleanup column kit. The final eluate is the cloning insert.

4. Perform restriction digestion to linearize the AAV vector at the unique cloning site to insert the enhancer into the vector.
    1. The cloning site in the 3’ UTR of the enhancer reporter construct used here is bounded by unique PacI and AscI restriction sites. Perform a digest to linearize the plasmid at this location according to the following parameters (steps 1.4.2–1.4.4).

2. Digest 1–10 µg of vector DNA using PacI and AscI, depending on how many cloning reactions will be needed. Prepare the reactions using the supplied buffer according to the manufacturer's instructions.
3. Incubate at 37 °C for 1–2 h to digest the vector DNA. (If digesting more than 5 µg in a 50 µL reaction, longer incubation may be necessary.)
4. After 1–2 h of incubation, inactivate the enzyme by incubating at 65 °C for 20 min.
5. Separate the restriction digest fragments by gel electrophoresis using a 0.7%–1% agarose gel, run at 100 V for 30–60 min.
6. Excise the band for the linearized vector and purify using a commercial gel extraction kit.

5. Perform Gibson cloning reactions and transform
    1. Determine the concentration of the purified vector and the cloning insert using a spectrophotometer.

       NOTE: DNA absorbs light at 260 nm, while contaminants absorb light at 230 nm and 280 nm. High ratios (>1.8) of 260/230 and 260/280 indicate highly purified DNA.
    2. Assemble Gibson cloning reactions in 0.2 mL PCR tubes: 5 µL of 2x Gibson mastermix, 0.02–0.5 pmol DNA fragments in a ratio of 3:1 insert to vector (Optimal DNA concentration for the reaction should be determined empirically) and Nuclease-free water (bring up the volume to 10 µL).
    3. Incubate Gibson reactions at 50 °C for 20 min in a thermal cycler.
6. Transform recombination deficient ($recA^-$) competent Escherichia coli (E. coli).
    1. Set a water bath to 42 °C and thaw the commercially available competent cells on ice.

NOTE: This step can be done prior to step 1.4, and the cells can thaw while preparing and incubating the Gibson reaction.

2. Aliquot 30–50 µL of cells into 2.0 mL microcentrifuge tubes. Add 2 µL of Gibson reaction to an aliquot and label the tube with the identity of the Gibson. Incubate the cells on ice for 30 min.

   NOTE: Use 5–10 ng of undigested empty vector as a positive control and water as a negative control.

3. Heat-shock the cells in the 42 °C water bath for 30–45 s, immediately return the tubes to ice, and let them recover for 5 min.

4. While on ice, add 400–950 µL of the commercially available recovery media.

   NOTE: The recovery media used in these experiments contains 2% vegetable peptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl2, 10 mM MgSO4, and 20 mM glucose.

5. After recovery on ice, fully recover the cells by incubating at 37 °C with gentle agitation of 250 RPM for 30–60 min.

6. Pellet the cells by centrifugation at 5,000 × g for 5 min and remove 400 µL of the supernatant, leaving ~50 µL of bacteria to plate.

7. Plate on selective media and incubate at 37 °C overnight.

   NOTE: The selective media used in these experiments contains 1.0% tryptone, 0.5% yeast extract, 1.0% sodium chloride, 1–2% agar, 96–97% water, and carbenicillin at a concentration of 100 µg/mL. Always use recombination deficient strains of bacteria for cloning viral plasmids because the viral inverted terminal repeats (ITRs) have matching sequences and can undergo homologous recombination. However, recA^{-} strains of bacteria still undergo low levels of recombination. The reporter

plasmid used here is a self-complementary AAV, where the sequence of one of the ITRs has been altered and no longer perfectly matches the other ITR. Growing the bacteria at 30 °C is also recommended for further slowing the rate of homologous recombination.

7. Verify clones and recover plasmids.
    1. Prepare a PCR master mix using forward and reverse primers (Table of Materials) that anneal to the vector, flanking the insert location. Aliquot 10 µL volumes into 0.2 mL PCR strip tubes.

Materials

Name	Company	Catalog Number	Comments
10× Citrate Buffer	Sigma-Aldrich	C9999-1000ML	
5'-gatcactctcggcatggac-3'	Integrated DNA Technologies	N/A: Custom designed	Forward primer for verifying clones after transformation. These primers are specific to the vector used and were designed for the specific vector used in our experiments.
5'-gatggctggcaactagaagg-3'	Integrated DNA Technologies	N/A: Custom designed	Reverse primer for verifying clones after transformation. These primers are specific to the vector used and were designed for the specific vector used in our experiments.
Agarose	VWR	VWRVN605-500G	
Aspirator tube assemblies	Sigma-Aldrich	A5177-5EA	for mouth-driven delivery of rAAV
Bacteriological petri dishes	Thermo Fisher Scientific	08-757-100D	
Carbenicillin	Sigma-Aldrich	C1389-5G	
Chicken IgY anti-GFP	Thermo Fisher Scientific	A10262	
Confocal microscope	Zeiss	LSM900	The images were taken on the LSM800 model, but Zeiss launched the LSM900 model in recent years to replace LSM800.
Conical centrifuge tubes 15 mL	Thermo Fisher Scientific	12-565-269	
Cryomolds	Thermo Fisher Scientific	NC9806558	These molds are suitable for P28 mouse brain. Other sizes may be more suitable for larger or smaller tissues.
DAPI	Sigma-Aldrich	D9542-10MG	
Dissecting scissors, 4.5"	VWR	82027-578	
Donkey anti-chicken AlexaFlour-488	Jackson ImmunoResearch	703-545-155	

Dulbecco's PBS 1×	Thermo Fisher Scientific	MT21031CV	
Eppendorf Microcentrifuge tubes 2.0 mL	Thermo Fisher Scientific	22431048	
Falcon round-bottom tubes 14 mL	Thermo Fisher Scientific	352059	
Fast Green dye	Grainger	F0099-1G	
Fine detail paint brush set	Artbrush Tower	B014GWCLFO	
Gibson Assembly Master Mix	NEB	E2611S	
Glass capillary tubes	Drummond Scientific Company	5-000-2005	
HiSpeed Plasmid Maxi Kit	QIAGEN	12663	Commercial plasmid maxi prep kit
HyClone HyPure Molecular Biology Grade Water	VWR	SH30538.03	
IV butterfly infusion set with 12" tubing and 25G needle	Thermo Fisher Scientific	26708	
Kimwipes	Kimberly Clark	34155	Lint-free wipe
LB Agar	Thermo Fisher Scientific	BP1425-500	LB agar pre-mix for selective media
McPherson Vannas iris scissor	Integra LifeSciences	360-215	
Mineral oil	Sigma Life Science	69794-500ML	
NEB Stable Competent E. coli	NEB	C3040I	
NucleoSpin Gel and PCR Clean-Up	Takara	740609.5	Kit for enzymatic reaction cleanup and gel extraction
OCT medium	VWR	25608-930	
Orbital shaker	Cole Parmer	60-100	
Paraformaldehyde	Sigma-Aldrich	158127-500G	
PCR strip tubes 0.2 mL	VWR	490003-692	
Peristaltic pump	Gilson	F155005	
Phosphate buffered saline (PBS) 10×	Thermo Fisher Scientific	70011044	
Phusion Hot Start II High Fidelity DNA Polymerase	Thermo Fisher Scientific	F549L	
Powdered milk	Sunny Select		
ProLong Gold Antifade Mountant	Thermo Fisher Scientific	P36934	
QIAquick PCR Purification Kit	QIAGEN	28106	
rCutSmart Buffer	NEB	B6004S	Buffer for restriction digest with PacI, AscI, and XmaI
Restriction enzyme: AscI	NEB	R0558L	
Restriction enzyme: PacI	NEB	R0547L	
Restriction enzyme: XmaI	NEB	R0180L	
SOC outgrowth medium	NEB	B0920S	Recovery medium after transformation
Sucrose (RNase/DNase free)	Millipore Sigma	033522.5KG	
TAE buffer	Apex	20-194	

Transfer tubing	Gilson	F1179941	For peristaltic pump
Triton X100	Sigma-Aldrich	X100-100ML	
Wizard Plus SV Minipreps DNA Purification System	Thermo Fisher Scientific	A1460	Plasmid mini prep kit

2. In 14 mL round-bottom tubes, prepare 5 mL aliquots of antibiotic selective LB broth, one for each colony being tested via PCR and negative control.
3. Use a sterile toothpick or pipette tip to pick an individual colony and roughly scrape the colony down into the bottom of one PCR tube.
4. When the colony has dissociated into the PCR master mix, drop the toothpick or tip into one of the LB aliquots and label the 14 mL tube with the Gibson reaction and the PCR tube number.
5. Place the colony PCR reactions on a thermal cycler and cycle with the program described in Table 2.

Table 2: Thermocycling conditions for colony PCR

Thermocycling conditions			
30 cycles of the following:			Notes
Step	Temperature	Time	
Cell Lysis	98 °C	5 min	
Denature	98 °C	10 s	
Anneal	55 °C	15 s	Optimal temperature may vary depending on primers.
Extend	72 °C	30 s	Optimal temperature and duration may vary depending on polymerase.
Hold	4 °C		

6. Incubate the 5 mL liquid cultures inoculated with toothpicks or pipette tips in a shaking incubator set to 37 °C and 180 RPM.
7. Run the colony PCR reactions on a 0.7%–1% agarose gel at 100 V for 30–60 min and visualize the bands in a gel doc imaging system.
8. Discard the 5 mL cultures that do not display the predicted PCR band.

9. For each successfully integrated clone, allow the 5 mL cultures to grow overnight and then perform a plasmid mini prep from 4.5 mL using a commercial kit. Reserve 0.5 mL of the liquid culture and save at −80 °C as a glycerol stock.
10. Confirm that the transgenes contained within the plasmids are free of undesirable mutations using Sanger sequencing[42].
11. After the reporter construct has been confirmed to be successfully cloned, use the saved glycerol stock to inoculate a 5 mL starter culture and grow for 8–16 h in a shaking incubator at 30 °C and 180 RPM.
12. Once the broth is cloudy, dilute the starter culture 1,000x in 250–300 mL of fresh selective LB broth and incubate at least overnight in a shaking incubator at 30 °C and 180 RPM. Then purify the plasmid using a commercial plasmid maxi kit.

    NOTE: Bacteria will grow more slowly at 30 °C compared to 37 °C, but this is the best temperature for slowing the rate of spontaneous recombination when growing large cultures.
13. Perform an XmaI digest to test for recombination of the ITRs using steps 1.4.2–1.4.5, substituting XmaI instead of PacI and AscI. Visualize the bands on a gel doc imaging system.

    NOTE: An rAAV plasmid with both ITRs present will yield two bands on a gel following this digest: one the length of the rAAV genome and the other the length of the rest of the plasmid backbone. If there is only one band, then the ITRs have undergone homologous recombination, and the rAAV plasmid will not be capable of packaging viral particles. The rAAV plasmid backbone described here has been designed so that XmaI sites only

occur in the ITRs. If the enhancer candidate insert also contains any XmaI sites, update the predicted band results and analysis accordingly.

2. Obtain packaged rAAV.
    1. Use packaged rAAV (professionally or in-house) for the experiments in this study.

        NOTE: There are many options for packaging an rAAV. An rAAV can be professionally packaged at a company or a university core facility. The rAAV can also be packaged in-house using different techniques that range in complexity and need for specialized equipment[43, 44]. The primary concern is obtaining a vector that is high quality and that the infectious titer has been determined to a high degree of certainty. Both professionally packaged and in-house packaged rAAVs were used for different experiments described in this study.
3. Intracranially inject rAAV-packaged plasmid into neonatal mice.
    1. Prepare rAAV mixture for delivery.
        1. If the intent is to compare the expression patterns between different enhancer reporter viruses, equalize the titers of all viruses to be compared by diluting all to match the least concentrated virus.
        2. Mix together a 2:1 or 3:1 ratio of enhancer reporter virus and constitutively expressed control reporter virus and add a small amount (0.06% final concentration) of Fast Green dye.

            NOTE: Fast Green is a dye that is used widely in injections of experimental animals, including with AAV, to visualize the injection site during and immediately after the procedure[45, 46, 47, 48]. While this is an important quality assurance step, it is possible that the addition of dye could interfere with transduction. Reconsidering the use of injection dye may be important for protocol optimization in certain cases. The

constitutively driven control virus is absolutely critical for comparing independent injections. Virus injections will vary in location and extent of transduction from animal to animal. The constitutive control will enable excluding failed injections from analysis and provide a standard for normalization of variation in virus delivery.

3. Break the tip of a pulled glass pipette made from a µL-graduated capillary tube by using it to gently pierce a delicate lint-free wipe. Then insert the pulled glass pipette into the aspirator assembly consisting of a device for applying pressure connected to 15-inch long rubber tubing ending with a rubber gasket for holding a micro-capillary pipette.

   NOTE: The “device for applying pressure” may be a syringe or a mouthpiece. If a syringe is to be used, a second person will be required to apply pressure through the syringe while the experimenter manipulates the pipette in one hand and the animal in the other. If using a mouthpiece, enabling the experimenter fine control over both the positions of the animal and pipette and the pressure applied to the syringe, extra care to avoid virus contamination is required.

4. Apply negative pressure through the aspirator assembly to draw a small volume (~0.2–0.5 µL) of mineral oil into the glass pipette.

   NOTE: This step is very important to provide a barrier to prevent aerosolized viral particles from contaminating the aspirator assembly.

5. Set the prepared aspirator assembly aside and keep the virus on ice until ready to inject.

2. Cryo-anesthetize the neonatal mice in a dry plastic dish partially submerged in ice until the cessation of the pedal withdrawal reflex (~5 min).

3. Use negative pressure through the aspirator assembly to draw the virus mixture into the pulled glass pipette until the meniscus passes the 2 µL tick mark.
4. Remove the mouse from the cold chamber and position it on the benchtop. Locate the bilateral injection sites midway between lambda and bregma and midway between the sagittal suture and each eye. Sanitize both areas with an alcohol wipe.
5. Use the pulled glass pipette to pierce the skull of the neonatal mice and gently apply positive pressure through the aspirator assembly as the needle enters the brain. Watch the meniscus of the virus mixture. Resistance to the pressure applied will decrease when the tip of the pipette reaches the lateral ventricle. Dispense 1 µL of the virus, withdraw the pipette, and repeat for the other lateral ventricle.
6. Place the mouse on a heating pad or a warming chamber to recover. Once awake and active, return the animal to its home cage.

4. Collect tissue and perform immunohistochemistry.
    1. After sufficient time following virus transduction, anesthetize the mice and sacrifice by transcardial perfusion.

        NOTE: Many experiments, including the representative results shown here, allow a period of 4 weeks or longer for the virus to transduce. However, self-complementary AAVs can show strong expression as early as 7 days[12]. The desired incubation period may need to be optimized depending on specific experimental techniques and goals.

        1. Prepare a peristaltic pump with intravenous infusion tubing and a 25 G needle.
        2. Prepare 1x PBS and 4% paraformaldehyde (PFA) and place them on ice.

3. Place the intake end of the tubing in the ice-cold 1x PBS. Set the flow rate (2.5 mL/min for P7 mice, 3.5 mL/min for P28 mice) and run the pump until the buffer is drawn all the way through the tubing and is cleared of bubbles. Set the needle aside until ready for perfusion.
4. Place the previously injected mouse in a drop jar chamber with a small volume of isoflurane and seal the chamber to anesthetize the mouse. Wait for ~5 min and then open the chamber and check for a pedal withdrawal reflex. Proceed once the animal is unconscious.

   NOTE: Throughout the perfusion, it is important to keep extra isoflurane on hand in case the mouse starts to regain consciousness. If this occurs, a small amount of isoflurane may be deposited on a delicate lint-free wipe and placed over the mouse's face for a few seconds until the animal is again fully unconscious.
5. Use surgical scissors to open the abdomen and chest of the mouse, remove the rib cage and expose the heart.
6. Use a pair of iris micro scissors to make a small incision in the mouse's right atrium, insert the IV needle into the left ventricle, and activate the peristaltic pump.
7. Perfuse the animal with ice-cold 1x PBS until the blood flushing out of the incision in the right atrium runs clear. Clearing the tissue of blood is very important because red blood cells have autofluorescence, and blood vessels may impair the ability to image reporter-expressing cells.
8. Pause the peristaltic pump and move the intake tubing from the PBS to the 4% PFA. Reactivate the pump and perfuse 1 mL/g of body weight, ~25 mL for an adult mouse.

NOTE: Fixation tremors should be observable, and the body of the mouse should stiffen.

2. Dissect and post-fix the brain.
    1. Remove the mouse's head with surgical scissors.
    2. Using small surgical scissors, start at the base of the skull and make an incision along the midline of the head and pull back the skin and underlying tissue to expose the skull.
    3. Using iris scissors, carefully cut the back of the skull, starting at the foramen magnum and continuing up toward and along the sagittal suture until the olfactory bulbs are passed.
    4. Insert the scissors into the skull over the olfactory bulbs and make two horizontal cuts in the skull. Make a similar pair of horizontal cuts in the occipital bone. The horizontal and midline cuts create a pair of door-like flaps in the top of the skull.
    5. Peel back the flaps of the skull to fully expose the brain. Use a pair of tweezers to sever the olfactory bulbs from the skull and to cut the cranial nerves, freeing the brain from the skull.
    6. Drop the brain into a 15 mL conical tube with 3–5 mL of 4% PFA in PBS. Post-fix 6 h to overnight. Do not over-fix the tissue.
3. Prepare the brains for cryosectioning.
    1. Transfer the fixed brain to a new 15 mL conical tube with 12–14 mL of 30% sucrose in PBS for dehydration. Incubate at 4 °C until the brain sinks to the bottom of the tube (2–3 days).
    2. Submerge the fixed and dehydrated brain in optimum cutting temperature (OCT) compound in a 15 mm × 15 mm × 5 mm cryomold. Add OCT until the brain is fully covered.

3. Place cryomold on a block of dry ice and freeze the sample in a solid block of OCT (~ 5 min).

   NOTE: OCT blocks can be stored at −80 °C for months to years.

4. Label the cryomold with the sample name and the orientation of the brain.

4. Obtain coronal sections using a cryostat microtome.

   1. Mark the orientation of the brain on the OCT block in pencil and remove the block from the cryomold inside the cryostat.
   2. Position the OCT block so that the brain is oriented with the olfactory bulbs facing the blade of the cryostat and adhere the block to the object holder of the cryostat with some fresh OCT.
   3. Once frozen in place, begin cutting 50 µm sections through the block following steps 4.4.4–4.4.6.
   4. Do not use the anti-roll plate. Sections will curl into tight rolls as they are cut and will collect either on top of the block or fall below into a collection tray.
   5. Watch the shape of the block in the first few cuts that are made. Make vertical cuts through the block, producing an even face when beginning to cut. Adjust the angle using the sample adjustment arms if needed.
   6. When the olfactory bulbs are visible, pay close attention to the shape of the sections. If one appears larger than the other, the cuts are at an angle relative to the brain, and it is necessary to straighten the orientation of the brain against the blade using the sample adjustment arms.

      NOTE: If cutting straight, the cortex should begin to appear above each olfactory bulb at the same time.

7. Once the olfactory bulb has been sectioned and rostral brain tissue is visible, reduce the cutting thickness to 30–35 μm and begin to collect the floating sections. Follow steps 4.4.8–4.4.13 for sectioning the brain.
8. Brush all previous sections off of the block and object holder so that they fall into the collecting tray. Brush them off to one side of the tray and keep them separate. If the pile is too big, then empty the tray.
9. Dispense 1 mL of PBS into each well of a 24 well plate. Label 1 row for each brain.
10. Cut 6 coronal sections. Retrieve the sections using cold tweezers and arrange them on the cryostat stage. Repeat this step 2 or 3 times, keeping the groups of 6 sections separate.
11. Wet a size 000 paintbrush with PBS and dab away excess liquid on a lint-free tissue or paper towel. Use the paintbrush to gently pick up each section, one at a time, and place it in the appropriate wells of the 24 well plate.
12. Place one of the 6 sections from the group of slices into each of the 6 wells in a row of the 24 well plate. This ensures that each well contains representative sections that span the sectioned area of the brain.
13. Continue sectioning in groups of 6 until the caudal end of the cerebral cortex has been sectioned.
14. For storage, seal the 24 well plate with parafilm and store at 4 °C.

    NOTE: Sections are stable for 1–2 months at 4 °C. A PBS solution with 0.1% Sodium azide may be used to inhibit bacterial growth and prolong the shelf life of the sections. However, if the sections are to be stored for longer than 2 months, they should be placed in cryoprotectant solution and frozen at −20 °C for best results.

5. Perform immunohistochemistry for reporter gene signal amplification.

   NOTE: All steps except antigen retrieval should be performed using gentle-to-moderate agitation (~100 RPM) on an orbital shaker.To preserve native fluorescence of the control reporter (mRuby3), all steps should be protected from light as much as possible.

   1. Prepare necessary buffers as described in Table 3.

Table 3: Buffers for immunohistochemistry

Buffers for immunohistochemistry	
Buffer	Composition
1x PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 1.8 mM KH2PO4
1x Citrate Antigen Retrieval Buffer (ARB), pH 6.0	Proprietary, see table of materials
Permeabilization Buffer (PB)	1x PBS + 0.5% Triton X100
Wash Buffer (WB)	1x PBS + 0.1% Triton X100
Blocking Buffer (BB)	WB + 5% milk

   2. Prepare a new 24 well plate with one row for each brain being stained. In the first column of wells, add 1 mL of PBS to each well.
   3. Using a size 000 paintbrush, gently transfer 1 well of sections from the original collection well to the fresh PBS in the new 24 well plate for a quick rinse to remove residual OCT compound.
   4. Add 1 mL of ARB to the next well in the 24 well plate and transfer sections to this well. Incubate in an oven at 60 °C for 1 h.
   5. Allow the plate to cool at room temperature (RT) for 20 min.
   6. Add 1 mL of PB to the next well and transfer the sections to this well. Incubate at RT for 20 min.
   7. Add 500 µL of BB to the next well and transfer the sections to this well. Incubate at RT for 1 h.

8. Add 2 mL of WB to the BB and then pipette out ~2 mL of solution and discard, leaving a diluted BB solution in the well. Repeat until the sections are clearly visible.
9. Dilute the primary antibody (Chicken IgY anti-GFP, 1:1000) in BB. Add 350–500 µL of primary antibody solution to the next well and transfer the sections to this well. Seal the plate with parafilm and incubate at 4 °C overnight.
10. Dilute BB with WB as done previously until the sections are clearly visible.
11. Add 1 mL of WB to the next well and transfer the sections to this well. Incubate at RT for 20 min.
12. Remove ~800–900 µL of buffer and discard. Add 1 mL fresh WB and incubate 20 min. Replace 1 mL of WB 4 more times for a total of 5 washes, 20 min each.
13. Dilute the secondary antibody (Donkey anti-chicken Alexa Fluor 488 conjugated, 1:500) in BB. Add 350–500 µL of the secondary antibody solution to a new well. Incubate at RT for 45 min to 1 h.
    NOTE: This is the seventh well, so if staining sections from more than 2 brains, a new plate will be needed at this point.
14. Dilute BB with WB as previously and transfer the sections to a new well filled with 1 mL of WB. Wash the sections as done previously for 20 min 3 times.
15. Prepare a working solution of DAPI in PBS (final concentration 0.04–0.8 µg/mL). Protect solution from light.
16. Add 1 mL DAPI solution to the next well and transfer the sections to this well. Incubate at RT for ~5 min.

17. Transfer the sections back to WB and gently mount on microscope slides using a size 000 paintbrush. Allow the slides to air dry.
18. Apply coverslips with an appropriate mounting media. Allow slides to cure at RT in the dark overnight.

5. Image and analyze brain tissue sections for enhancer activity.
    1. Use a low magnification (5x) objective lens to record fluorescence images that span the brain section.
    2. Locate the injection site and evaluate the extent of the virus transduction based on the density and intensity of red fluorescent cells. Take care to compare animals that were transduced similarly.
    3. Record detailed fluorescent images with a higher magnification objective lens (≥25x) if desired.

        NOTE: Always be sure to use the same settings for acquisition parameters such as excitation light intensity, filters, and exposure time between samples that are to be compared.
    4. Process the images using image analysis software.

        NOTE: Processing may be done in any image analysis software package. The following steps are suggestions for determining whether a sequence acts as an enhancer in the reporter construct context (a yes or no question) using the open-source FIJI distribution of ImageJ. More in-depth photometry may be appropriate when comparing degrees of activity between enhancer variants. Images from different samples should always be processed consistently in order to be compared.
        1. Apply filters to reduce noise. In Fiji select the Despeckle option under the Process > Noise menu.

            NOTE: This is a 3 × 3 median filter.

2. Subtract background fluorescence following steps 5.4.3–5.4.6.
3. Separate the channels of a multichannel image. In Fiji, select the option Split Channels under the Image > Color menu.
4. Use the rectangle or circle selection tool to draw a small shape in an area of the image that does not have any fluorescent cells and measure the average pixel intensity of the background using Analyse > Measure. Repeat to sample multiple areas.

   NOTE: Make sure Mean Gray Value is selected in the Analyze > Set Measurements menu.
5. Average the mean gray value from 5–8 areas of background and drop the decimal point. Subtract this value from each pixel in the image using Process > Math > Subtract.

   NOTE: Rounding to the nearest whole number could over-estimate the background to subtract. It is more conservative to simply drop the decimal point. For example, 6.4 would be rounded down to 6, but 6.5 should also be rounded down to 6 to avoid the risk of subtracting 0.5 units of real signal from each pixel.
6. If desired, merge the channels back into a multichannel image. In Fiji, use Image > Color > Merge Channels.
7. Stitch any overlapping images together. In Fiji, use Plugins > Stitching > Grid/Collection stitching.
8. Adjust brightness and contrast. In Fiji, use Image > Adjust > Brightness/Contrast.

   NOTE: Always use the same minimum and maximum values for each image to be compared.

9. Count the number of green fluorescent cells following steps 5.4.10–5.4.12. These are the cells in which the candidate enhancer sequence was able to drive reporter expression.
10. Start by hiding the green channel and use the freeform selection tool to draw a shape around the region of the brain that expresses red fluorescence. Using the Measure function in Fiji, measure the area, integrated density, and mean gray value of the red channel in this zone.
11. Using the multi-point selection tool, count the number of green cells in this zone.
12. Normalize the number of green cells by the integrated density of the red channel. The local red channel brightness is a proxy measure for the magnitude of virus exposure, enabling the comparison of enhancer activity between different transduced animals.

Representative Results

Using these methods, a 915 bp sequence in the psychiatric risk-associated third intron of the gene *CACNA1C*[19, 49, 50] was tested for enhancer activity in the postnatal mouse brain. This sequence was discovered in an MPRA of 345 candidate enhancer sequences centered on psychiatric and neurological risk SNPs[12] and characterization experiments are described here as a general example. C57BL/6 mice were injected at P0 with an AAV9 construct packaged as a self-complementary vector[51] containing EGFP under the control of the Hsp68 minimal promoter and a single candidate enhancer sequence (Figure 3A). Additional mice were injected with constructs that contained a negative control sequence that was predicted to have no regulatory activity (Figure 3B). Viral titers for these constructs were normalized to 6.85×10^{10} vg/mL. All injected mice were co-injected with an AAV9-packaged construct containing mRuby3 under the control of the constitutive CAG promoter to visualize the area that was successfully

transduced and control for potential variability in the site and spread of infection. Brains were collected at postnatal day (P)28 for immunohistochemistry and imaging. These experiments confirmed that this candidate enhancer region in *CACNA1C* drove expression of EGFP in P28 mouse brain (Figure 3A), while a negative control sequence did not (Figure 3B). These results demonstrate the application of enhancer-reporter assays in the mouse brain, enabling *in vivo* study of enhancers during conditions that cannot be recapitulated using traditional *in vitro* models.

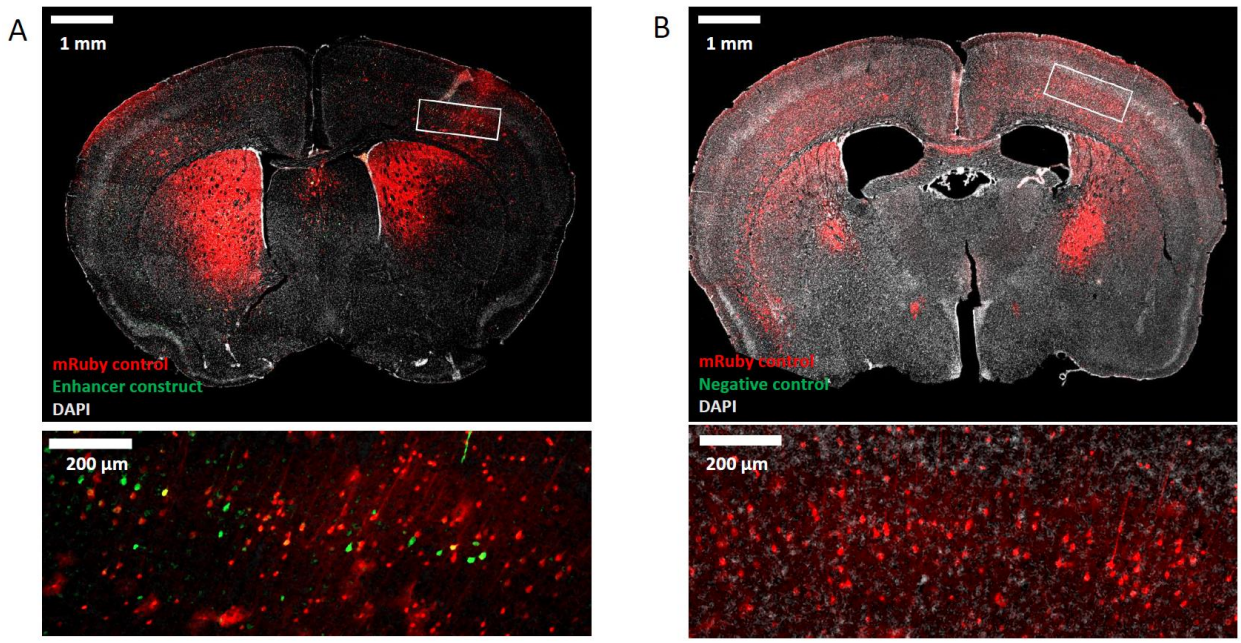


Figure 3: Enhancer in *CACNA1C* drives EGFP expression in P28 mouse brain.

(A) A mouse injected intracranially at P0 with an enhancer-reporter construct (scAAV9-Hsp68-EGFP-*CACNA1C*_3) and constitutively expressed injection control (AAV9-CAG-mRuby3) shows enhancer-driven EGFP expression in the middle-lower layers of the cortex at P28. (B) A mouse injected at P0 with a negative control construct (scAAV9-Hsp68-EGFP-NEG_4) and injection control does not show expression of EGFP at P28. Whole-brain images were taken at 5x magnification and insets show zoomed-in view of the boxed region.

These methods can also be used to test libraries of candidate enhancer sequences for activity using AAV-based delivery methods, such as in MPRAs. Representative images of library-based reporter expression at different titers demonstrate the flexibility of this assay and the effect of viral titer on transduction efficiency. High-titer (Figure 4A) versus low-titer (Figure 4B) viral preparations result in differences in the amount of transduced cells, as evidenced by both positive control mRuby3 expression driven by a CAG promoter and EGFP expression produced from the candidate enhancer libraries in postnatal mouse brain. There are situations where either high titer and broad transduction or low titer and sparse transduction may be preferred.

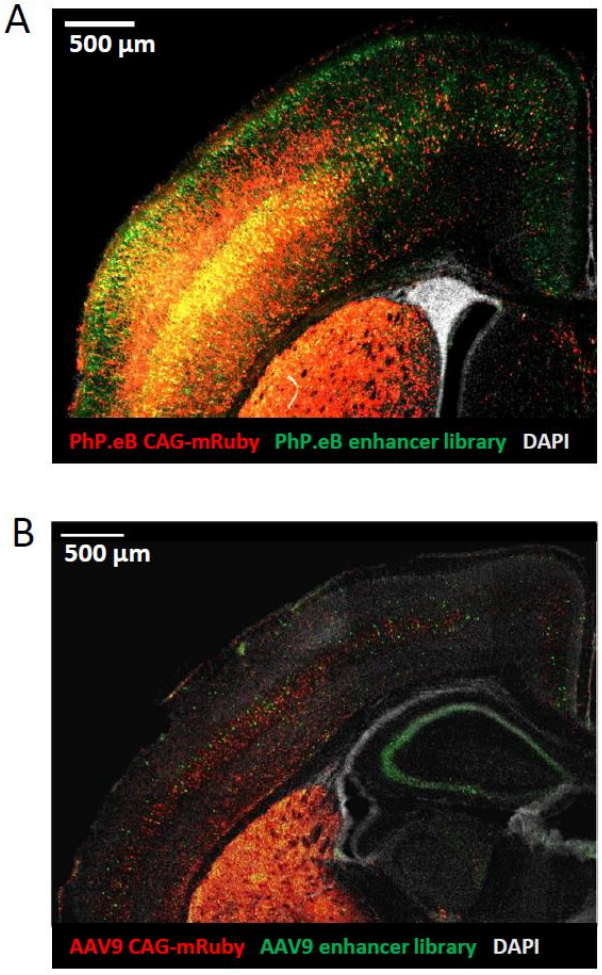


Figure 4: AAV-delivered enhancer-reporter libraries of different titers show activity in mouse brain.

(A) Commercially packaged, high-titer AAV-PHP.eB library of candidate enhancers drives broad EGFP expression in P7 mouse brain. (B) Lower-titer AAV9 library of candidate enhancers precipitated from conditioned media of the packaging cells drives sparse EGFP expression in P7 mouse brain. Images were taken at 5x magnification.

Discussion

This protocol describes an rAAV-based method for the deployment of enhancer-driven transgenes in the postnatal mouse brain. In this generalized protocol, a candidate enhancer, a minimal promoter, a reporter gene, and an optional barcode sequence are cloned into an AAV plasmid backbone. These experiments can be done with a single candidate enhancer sequence or with many sequences in parallel. The plasmid is packaged into an rAAV and delivered to the postnatal mouse brain. After a period of time to allow for virus transduction, the brain is collected and imaged for reporter gene expression. A constitutively expressed reporter virus (CAG-mRuby3) is included as an injection control. Using this assay, enhancer elements are shown to drive transcription of EGFP in the mouse brain, demonstrating enhancer activity *in vivo*.

The primary targets for troubleshooting and optimization of this assay include viral titer, injection technique, and transduction time frame. Different viral titers can have a strong effect on the density of transduced cells. While a higher transduction efficiency is likely preferable for most research, the optimal level of transduction is highly dependent on experimental goals. Intracranial injections offer a high density of transduced cells, but are likely to be more focal at the site of injection, though the spread of infection may depend on viral serotype and age of the animal[52]. Retro-orbital injections may show more even distribution of transduced cells throughout the brain but are less likely to give any area of high infection. Tail-vein injections can more effectively transduce other tissues in addition to the brain but will likely give still lower densities of transduced cells in the brain. Similarly, a variety of time-frames for virus transduction in the brain can serve different purposes. Robust scAAV expression can occur in the brain as early as seven days after intracranial injection, though periods of 28 days or longer are more typically used and yield strong expression. As enhancers can be temporally specific in

their activity, it is important to consider the predicted activity of the tested enhancers when determining study design.

There are some limitations associated with these methods that might make other options a more suitable choice, depending on experimental goals. Since AAVs exhibit low integration into the genome, tissues are most effectively transduced when they are mature and there is a high density of postmitotic cells, as a large amount of continued cell divisions will lower the density of transgene-expressing cells. For viral delivery of a transgene to a less mature model, such as tissues that are still growing, lentiviruses may be a more appropriate choice, as they integrate into the genome and will be inherited by all daughter cells of a transduced proliferating cell. Additionally, while the reporter assay described here is a widely used technique for the functional study of *cis*-regulatory activity, it cannot provide a complete picture of how the regulatory element acts within its native context. For example, this assay does not provide information on what genes may be targeted by the regulatory element in the genome. However, if the enhancer's target genes are known, this information can be leveraged to determine if the cell types showing expression in this assay are the same as those known to express the enhancer's targeted genes, which would lend high biological validity to the interpretation of results. The assay also may not completely recapitulate the effects that the surrounding genomic sequence or normal biological and behavioral activity of the animal may have on the activity of the tested regulatory element. Finally, this protocol describes the STARR-seq reporter assay design, in which the candidate enhancer element is cloned downstream of the reporter gene in the ORF, in contrast to upstream of the promoter, as in the conventional reporter assay cloning orientation. Including long variable sequences in the ORF of the reporter gene may cause reduced RNA stability due to factors such as RNA binding protein motifs, alternate splice sites, or variable GC content [36, 53], and statistical methods have been developed to account for sequence-based biases when analyzing STARR-seq data[54, 55]. Despite these considerations,

STARR-seq MPRAs have been widely used in recent years to successfully identify *cis*-regulatory elements[56, 57, 58, 59]. The methods described here can be easily adapted for use with the conventional MPRA orientation.

The cell-type specificity of enhancers makes this assay a powerful tool in driving the cell-type-specific expression of genes of interest *in vivo*. Current techniques for cell-type-specific expression of a desired gene can be limited by the availability of conditional expression constructs or require the generation of new transgenic mouse lines. Using an rAAV-based enhancer-driven system, a gene of interest can be expressed in a spatiotemporally specific manner by using validated cell-type-specific enhancers or promoters to drive expression[13, 14, 15]. Recent developments in this technology have shown that combining cell-type-specific enhancers with rAAV serotypes with tropisms for cells or tissues of interest can show similar specificity of transgene expression as in transgenic animal models[60]. This offers a cheap, fast, and potentially high-throughput method for sequence screening and specific induction of transgene expression *in vivo* in animal models.

This technology also has therapeutic potential. Clinical studies have found that rAAV-based delivery of a transgene *in vivo* can induce the expression of healthy copies of a gene in cases where only a defective copy is present, or the gene is not being transcribed at a sufficient rate[61, 62, 63]. The selection of a driving enhancer element has the potential to enable generalized delivery but expression induction in a cell-type-specific manner. Finally, delivering the transgene via rAAV also confers significant benefits; AAVs have lower immunogenicity than other viruses often used in transgene delivery[64], making this a potentially safe viral vector to be employed for clinical use.

Specific deployment of *in vivo* rAAV-based enhancer-reporter assays is customizable and can be modified to suit a range of experimental goals. Different tissues and cells can be targeted by

selecting AAV serotypes or enhancers that have cell type or spatiotemporal specificity. Anywhere from one to thousands of enhancer-reporter constructs can be delivered *in vivo*, and activity can be assayed using a number of imaging- and sequencing-based readouts, as have been shown in an emerging set of studies using this approach[15, 16, 65, 66, 67, 68]. The range of options for construct design and delivery enables this technique to be modified for a variety of uses in both translational and basic science, making this a powerful new tool for genomics and neuroscience research.

Acknowledgments

Sequencing was performed at the UC Davis DNA Technologies Core. We thank the lab of Lin Tian at UC Davis for training on rAAV packaging and generously gifting us AAV helper and rep/cap plasmids. This work was supported by NIH/NIGMS R35GM119831.

Footnotes

A complete version of this article that includes the video component is available at http://dx.doi.org/10.3791/62650.

Disclosures

The authors declare no competing financial interests.

References

1. Pennacchio LA et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 444 (7118), 499–502 (2006).
2. Nord AS et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell. 155 (7), 1521–1531 (2013).
3. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics. 47 (11), 1228–1235 (2015).

4. Nord AS, West AE Neurobiological functions of transcriptional enhancers. Nature Neuroscience. 23 (1), 5–14 (2020).
5. Hallikas O et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell. 124 (1), 47–59 (2006).
6. Su J, Teichmann SA, Down TA Assessing computational methods of *cis*-regulatory module prediction. PLoS Computational Biology. 6 (12), e1001020 (2010).
7. Visel A et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 457 (7231), 854–858 (2009).
8. Kvon EZ et al. Genome-scale functional characterization of Drosophila developmental enhancers *in vivo*. Nature. 512 (7512), 91–95 (2014).
9. Yáñez-Cuna JO et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Research. 24 (7), 1147–1156 (2014).
10. Zincarelli C, Soltys S, Rengo G, Rabinowitz JE Analysis of AAV serotypes 1–9 mediated gene expression and tropism in mice after systemic injection. Molecular Therapy. 16 (6), 1073–1080 (2008).
11. Gisselbrecht SS et al. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nature Methods. 10 (8), 774–780 (2013).
12. Lambert JT et al. Parallel functional testing identifies enhancers active in early postnatal mouse brain. eLife. 10, e69479 (2021).
13. Chen Y-JJ et al. Use of “MGE Enhancers” for labeling and selection of embryonic stem cell-derived medial ganglionic eminence (MGE) progenitors and neurons. PLoS ONE. 8, e61956 (2013).
14. Graybuck LT et al. Enhancer viruses and a transgenic platform for combinatorial cell subclass-specific labeling. bioRxiv. bioRxiv 525014 (2019).
15. Dimidschstein J et al. A viral strategy for targeting and manipulating interneurons across vertebrate species. Nature Neuroscience. 19 (12), 1743–1749 (2016).
16. Rubin AN et al. Regulatory elements inserted into AAVs confer preferential activity in cortical interneurons. eNeuro. 7, ENEURO.0211–20.2020 (2020).
17. Hindorff LA et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 106 (23), 9362–9367 (2009).
18. Maurano MT et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 337 (6099), 1190–1195 (2012).
19. Ripke S et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 511 (7510), 421–427 (2014).
20. Nott A et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science. 366 (6469), 1134–1139 (2019).
21. Lagunas T et al. A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of genetic variants *in vivo*. bioRxiv. biorxiv2021.05.17.444514 (2021).
22. DeBoer E, Antoniou M, Mignotte V, Wall L, Grosveld F The human beta-globin promoter; nuclear protein factors and erythroid specific induction of transcription. The EMBO journal. 7 (13), 4203–4212 (1988).

23. Kothary R et al. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. Development. 105 (4), 707–714 (1989).
24. Niwa H, Yamamura K, Miyazaki J Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene. 108 (2), 193–199 (1991).
25. Kim DW, Uetsuki T, Kaziro Y, Yamaguchi N, Sugano S Use of the human elongation factor 1α promoter as a versatile and efficient expression system. Gene. 91 (2), 217–223 (1990).
26. Banerji J, Rusconi S, Schaffner W Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell. 27 (2 Pt 1), 299–308 (1981).
27. Patwardhan RP et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nature Biotechnology. 27 (12), 1173–1175 (2009).
28. Arnold CD et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 339 (6123), 1074–1077 (2013).
29. Melnikov A et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature Biotechnology. 30, 271–277 (2012).
30. Kim J-Y, Grunke SD, Levites Y, Golde TE, Jankowsky JL Intracerebroventricular viral injection of the neonatal mouse brain for persistent and widespread neuronal transduction. Journal of Visualized Experiments: JoVE. 91, 51863 (2014).
31. Liu D, Zhu M, Zhang Y, Diao Y Crossing the blood-brain barrier with AAV vectors. Metabolic Brain Disease. 36 (1), 45–52 (2021).
32. Prabhakar S, Lule S, DA Hora CC, Breakefield XO, Cheah PS AAV9 transduction mediated by systemic delivery of vector via retro-orbital injection in newborn, neonatal and juvenile mice. Experimental Animals. (2021).
33. Gruntman AM, Su L, Flotte TR Retro-orbital venous sinus delivery of rAAV9 mediates high-level transduction of brain and retina compared with temporal vein delivery in neonatal mouse pups. Human Gene Therapy. 28 (3), 228–230 (2017).
34. Inoue F, Ahituv N Decoding enhancers using massively parallel reporter assays. Genomics. 106 (3), 159–164 (2015).
35. Gordon MG et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. Nature Protocols. 15 (8), 2387–2412 (2020).
36. Klein JC et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nature Methods. 17 (11), 1083–1091 (2020).
37. Creyghton MP et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America. 107 (50), 21931–21936 (2010).
38. Song L, Crawford GE DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols. 2010. (2), pdb.prot5384 (2010).
39. King DM et al. Synthetic and genomic regulatory elements reveal aspects of *Cis*-regulatory grammar in mouse embryonic stem cells. eLife. 9, e41279 (2020).
40. Amit R, Garcia HG, Phillips R, Fraser SE Building enhancers from the ground up: a synthetic biology approach. Cell. 146, 105–118 (2011).
41. Fan M, Tsai J, Chen B, Fan K, LaBaer J A central repository for published plasmids. Science. 307, 1877 (2005).

42. Sanger F, Coulson AR A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 94, 441–448 (1975).
43. Broussard GJ, Unger EK, Liang R, McGrew BP, Tian L Imaging glutamate with genetically encoded fluorescent sensors. in Biochemical Approaches for Glutamatergic Neurotransmission. (eds. Parrot S, & Denoroy L). 130, 117–153, Springer, New York: (2018).
44. Chen YH, Keiser MS, Davidson BL Adeno-associated virus production, purification, and titering. Current Protocols in Mouse Biology. 8, e56 (2018).
45. He CX, Arroyo ED, Cantu DA, Goel A, Portera-Cailliau C A Versatile Method for Viral Transfection of Calcium Indicators in the Neonatal Mouse Brain. Frontiers in Neural Circuits. 12, 56 (2018).
46. Chen S-Y, Kuo H-Y, Liu F-C Stereotaxic surgery for genetic manipulation in striatal cells of neonatal mouse brains. Journal of Visualized Experiments: JoVE. 137, e57270 (2018).
47. Brenowitz SD, Regehr WG Presynaptic imaging of projection fibers by *in vivo* injection of dextran-conjugated calcium indicators. Cold Spring Harbor Protocols. 2012, pdb.prot068551 (2012).
48. Bosch D, Asede D, Ehrlich I Ex vivo optogenetic dissection of fear circuits in brain slices. Journal of Visualized Experiments: JoVE. 110, e53628 (2016).
49. Green EK et al. The bipolar disorder risk allele at *CACNA1C* also confers risk of recurrent major depression and of schizophrenia. Molecular Psychiatry. 15 (10), 1016–1022 (2010).
50. Takahashi S, Glatt SJ, Uchiyama M, Faraone SV, Tsuang MT Meta-analysis of data from the Psychiatric Genomics Consortium and additional samples supports association of *CACNA1C* with risk for schizophrenia. Schizophrenia Research. 168 (1–2), 429–433 (2015).
51. McCarty DM Self-complementary AAV vectors; advances and applications. Molecular Therapy. 16 (10), 1648–1656 (2008).
52. Haery L et al. Adeno-associated virus technologies and methods for targeted neuronal manipulation. Frontiers in Neuroanatomy. 13, 93 (2019).
53. Rabani M, Pieper L, Chew G-L, Schier AF A massively parallel reporter assay of 3′ UTR sequences identifies *in vivo* rules for mrna degradation. Molecular Cell. 68, 1083–1094.e5 (2017).
54. Kim Y-S et al. Correcting signal biases and detecting regulatory elements in STARR-seq data. Genome Research. 31, 877–889 (2021).
55. Lee D et al. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. Genome Biology. 21, 298 (2020).
56. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD Candidate silencer elements for the human and mouse genomes. Nature Communications. 11, 1061 (2020).
57. Singh G et al. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. Genome Research. 31, 564–575 (2021).
58. Wang X et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nature Communications. 9, 5380 (2018).
59. Peng T et al. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. Genome Biology. 21, 243 (2020).

60. Nair RR, Blankvoort S, Lagartos MJ, Kentros C Enhancer-driven gene expression (EDGE) enables the generation of viral vectors specific to neuronal subtypes. iScience. 23, 100888 (2020).
61. Russell S et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. The Lancet. 390, 849–860 (2017).
62. Gaudet D et al. Efficacy and long-term safety of alipogene tiparvovec (AAV1-LPLS447X) gene therapy for lipoprotein lipase deficiency: an open-label trial. Gene Therapy. 20, 361–369 (2013).
63. Nguyen GN et al. A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. Nature Biotechnology. 39, 47–55 (2021).
64. Nayak S, Herzog RW Progress and prospects: Immune responses to viral vectors. Gene Therapy. 17 (3), 295–304 (2010).
65. Visel A et al. A high-resolution enhancer atlas of the developing telencephalon. Cell. 152 (4), 895–908 (2013).
66. Shen SQ et al. Massively parallel *cis*-regulatory analysis in the mammalian central nervous system. Genome Research. 26 (2), 238–255 (2016).
67. Shen SQ et al. A candidate causal variant underlying both higher intelligence and increased risk of bipolar disorder. bioRxiv. bioRxiv10.1101/580258 (2019).
68. Hrvatin S et al. A scalable platform for the development of cell-type-specific viral drivers. eLife. 8, e48089 (2019).

c. Appendix

Validation of enhancer activity in the schizophrenia gene *CACNA1C*

An MPRA previously published by our lab identified three enhancers in the schizophrenia-associated third intron of the gene *CACNA1C* as active in postnatal mouse brain[1] (Fig. 1A). I conducted experiments during my dissertation to validate the activity of these elements. Given the importance of adolescent neurodevelopment to schizophrenia, my first goal was to extend the time course of our validation work on *CACNA1C* to verify activity in adolescent mouse brain. I used the methods described in the above manuscript to test for activity at *CACNA1C* enhancer candidate #3 using an AAV-based assay in postnatal mouse brain. For this experiment, I injected my AAV containing the *CACNA1C* #3 enhancer-reporter construct

into P0 mice. I also co-injected a positive control AAV containing mRuby3 under the control of the constitutive promoter CAG. I collected brains at P28 for imaging for EGFP. I found strong expression of EGFP and mRuby3 in P28 mouse brain (Fig. 1B). This *in vivo* work is featured in Figure 4 and its second corresponding supplementary figure in our lab's paper describing this MPRA, on which I am fourth author[1]. In addition to validating *in vivo* activity of *CACNA1C* #3, I extended our lab's MPRA abilities into testing activity in adolescent mouse brain, which forms much of the basis of the second chapter of my dissertation work.

My next goal for this work was to test for enhancer activity in HEK-293T cells in order to validate their activity in a human cell line. I separately cloned our three top enhancers in *CACNA1C* into our lab's MPRA plasmid backbone: *CACNA1C* #3, *CACNA1C* #6, and *CACNA1C* #7. I transfected these plasmids into separate wells of HEK-293T cells using Lipofectamine 3000 (ThermoFisher # L3000001). For controls, I transfected wells with empty plasmid without any regulatory element cloned into it ("empty vector"), as well as plasmids into which I cloned two candidate enhancer elements in *CACNA1C* which had not shown activity in our assay: *CACNA1C* #2 and *CACNA1C* #5. All candidates were transfected into individual wells at three wells each, such that 18 samples were transfected in total (three wells for each of three enhancers, two negative elements, and one empty plasmid). After 48 hours of incubation, I extracted RNA from the cells using the Qiagen RNeasy kit (Qiagen # 74104), including the on-column DNAse treatment with the Qiagen RNAse-Free DNase kit (Qiagen # 79254). I reverse transcribed the RNA into cDNA using SuperScript III and random hexamers (ThermoFisher # 18080400). Using Applied Biosystems *Power* SYBR Green PCR Master Mix (ThermoFisher # 4368577), I performed qPCR for *EGFP* in the cDNA obtained from all wells, as well as for the housekeeping gene GAPDH. I used the $2^{-\Delta\Delta Ct}$ method[2] to calculate the fold change in EGFP expression versus the empty vector control. I then conducted an ANOVA to test whether the conditions showed differences in activity, followed by Tukey's HSD test[3] to compare each

experimental group to the empty vector control. My results replicated our findings from the *in vivo* MPRA experiments. I identified significantly increased *EGFP* RNA for our three top enhancer candidates versus empty vector, while the two negative candidates from our MPRA did not show increases in *EGFP* RNA (Fig. 1C).

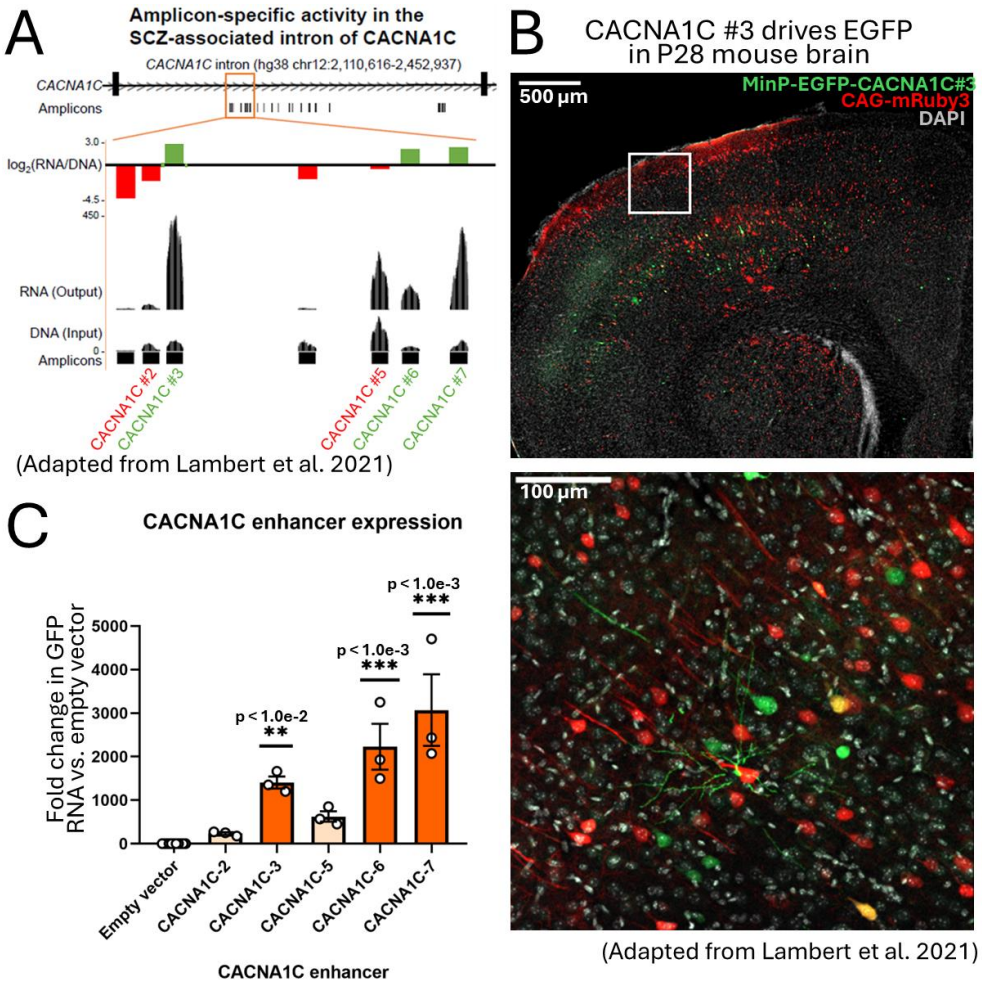


Fig. 1: Validation of MPRA-identified enhancers in the psychiatric risk-associated third intron of *CACNA1C*. (A) Our lab's published MPRA identified three enhancers in the third intron of *CACNA1C* which were active in P7 mouse brain (Adapted from Lambert et al. 2021)[1]. (B) I injected an AAV containing the *CACNA1C* #3 enhancer-reporter construct into P0 mouse brain using intracerebroventricular injections. I also co-injected a positive control AAV containing the mRuby3 reporter gene under the control of

the constitutive promoter CAG. Brains were collected at P28. The *CACNA1C* #3 enhancer-reporter construct, as well as the CAG-mRuby3 positive control, drove strong expression of their reporter genes (Adapted from Lambert et al. 2021)[1]. (C) I transfected enhancer-reporter plasmids containing these three elements, as well as two elements which had not shown activity and one plasmid which did not have any enhancer cloned into it ("empty vector"), into HEK-293T cells. qPCR showed that the three positive enhancers drove significantly increased *EGFP* RNA relative to empty vector, while the two negative elements did not.

References

1. Lambert, J. T.* & Su-Feher, L.* *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
* indicates co-first authors
2. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**, 402–408 (2001).
3. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* **5**, 99–114 (1949).

Chapter 2:
Schizophrenia-associated risk variants alter the function of regulatory sequences that show specific activity during adolescent brain development

a. Preface

The MPRA designs described in Chapter 1 constituted state-of-the-art technology at the time I joined our lab. As MPRAs have matured, the field has learned major lessons about how to improve their efficiency and scientific rigor. In response to new insights into MPRA technologies and following the work described in Chapter 1, I proceeded to restructure major aspects of MPRA design in our lab to better control against potential confounds and to streamline and modularize overall workflow. I subsequently deployed an optimized MPRA strategy to assess activity at candidate *cis*-regulatory elements (CREs) containing a set of prioritized schizophrenia-associated SNPs in postnatal mouse brain.

In this chapter, I first describe some of the challenges facing MPRA experiments and major design changes I made to our lab's MPRA protocols to address these considerations. These changes include transitioning our initial MPRA plasmid from the STARR-seq conformation to the conventional MPRA plasmid conformation, developing a highly complex library of plasmids containing MPRA barcode regions that were optimized to reduce the introduction of confounds, and designing and validating a library of negative control plasmids for use in MPRA experiments. I additionally extended our lab's capabilities for *in vivo* CRE testing from early postnatal to late adolescent time points, allowing us to further take advantage of our *in vivo* MPRA model and assess CRE activity in additional developmental contexts.

I then describe the deployment of this optimized, redesigned MPRA protocol to assess CRE activity of prioritized schizophrenia-associated risk loci in postnatal mouse brain. In addition to testing overall CRE activity at loci, I tested for the impact of schizophrenia-associated sequence

variation within these regulatory elements on transcription rate. I also assessed the developmental time course of CRE activity from juvenile to adolescent mouse brain. Finally, I intersected my findings with published epigenetic and gene expression datasets to generate hypotheses for gene targets for future testing. My results identified 39 active CREs in postnatal mouse brain, several of which show differential activity between juvenile and adolescent mice. Of active CREs, 19 show allelic differences. I highlight three loci for detailed analyses which showed allelic differences and adolescent-specific increases in activity. I independently conceived of, designed, executed, analyzed, and wrote the experiments described in this chapter.

b. MPRA protocol optimization

STARR-seq versus barcode-based design

The MPRA is a young and rapidly evolving experimental protocol, with a range of offshoots and iterations that have been developed in response to different experimental priorities. A shared feature among different designs is the presence of unique sequences in the RNA readout which are assigned to specific candidate regulatory elements being tested. This allows a sequenced RNA molecule to be confidently assigned to the CRE which drove its transcription. In the conventional MPRA configuration, these sequences are short, random nucleotide sequences that are cloned downstream of the reporter gene during the plasmid cloning process, known as barcodes, while CREs are cloned upstream of the minimal promoter[1,2] (Fig. 1A). Researchers then keep a key which details which CREs are mapped to which barcodes. This key is referenced during analysis of RNA sequencing data to understand which CRE is represented by a given barcode in the sequencing results. An update to the conventional MPRA design is known as STARR-seq (**S**elf-**T**ranscribing **A**ctive **R**egulatory **R**egion **seq**uencing), in which the CRE itself is cloned into the open reading frame, typically in the 3' untranslated region (3'UTR). The CRE sequence can subsequently be identified in the RNA readout and serve as its own

barcode[3] (Fig. 1B). This design takes advantage of the fact that, operating in the three-dimensional chromatin space, enhancers and silencers act agnostically to whether they are located upstream or downstream from their targeted gene promoter[4]. As such, CREs are able to initiate transcription at their own promoters. This phenomenon is well characterized in the native transcriptional activity of cells[5–7] and has been used repeatedly and effectively in MPRAs that employ the STARR-seq design[3,8,9].

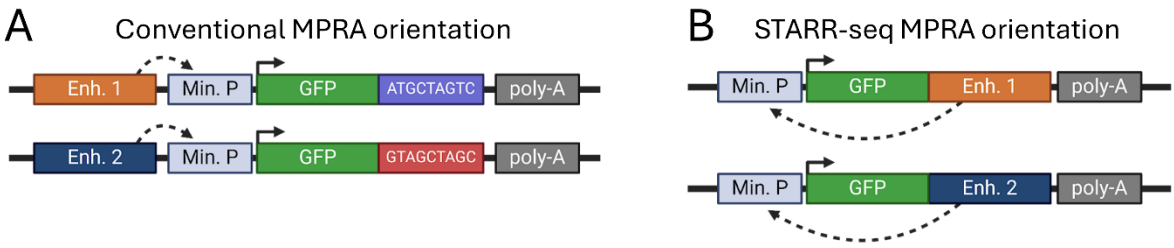


Fig. 1: Conventional versus STARR-seq orientation for massively parallel reporter assays. (A) In the conventional design, the candidate CRE is cloned upstream of the minimal promoter. One or more unique barcodes per CRE are cloned downstream of the reporter gene in the open reading frame, typically in the 3'UTR. These barcodes will be identified in the RNA sequencing reads and associated back to the CRE which drove their expression. (B) In the STARR-seq design, the candidate CRE is cloned downstream of the reporter gene in the open reading frame, typically in the 3'UTR. Active CREs will drive transcription off of their own minimal promoters. The CRE sequence itself will be identifiable in the RNA sequencing readout.

Our lab's initial MPRA design featured a plasmid in the then cutting-edge STARR-seq conformation. This design offers major advantages prior to launching an experiment. While inserting CREs and barcodes into separate positions, as in the conventional design, can require multiple cloning steps and some method to determine which CREs are associated to which barcodes, using a STARR-seq design streamlines cloning into a single step and eliminates the

need to determine CRE-barcode associations. The result is faster cloning and experimental launch, and a quick and easy understanding from RNA sequencing results of which CREs are being tested.

While the experimental setup is simpler and more streamlined, STARR-seq can also be subject to certain weaknesses that are mitigated in the conventional MPRA orientation. First, placing a long sequence of variable length and GC content in the 3'UTR may have unpredictable effects on RNA stability. The 3'UTR functions in large part to regulate RNA stability via post-transcriptional modifications[10]. Inserting large regulatory elements into this region may disrupt this process without leaving an opportunity to dissect whether apparent changes in RNA reads are due to CRE modification of transcription initiation, or due to 3'UTR impacts on RNA stability. Similarly, variable lengths and GC content of CREs cloned into the open reading frame may impact PCR efficiency when processing RNA and cDNA after completion of these experiments. In fact, GC content was associated with and subsequently controlled for in our models of enhancer activity in a previous MPRA publication from our lab[8]. This may have been due to altered PCR efficiency, altered RNA stability, or both in our experiments. Finally, using a CRE as its own barcode eliminates the possibility to assess technical reproducibility for a single CRE element within the same biological sample. RNA readouts from STARR-seq assays do not provide information as to whether one molecule or many molecules of the MPRA construct drove transcription for a given CRE. This may reduce confidence as to whether RNA data represents true regulatory activity or merely a statistical anomaly which may not actually be reproducible. Fortunately for experiments that employ STARR-seq, a systematic analysis of nine different potential MPRA plasmid designs showed moderately strong correlations of enhancer activity, averaged across biological replicates, between experiments using STARR-seq and other MPRA strategies[11]. Pearson's R^2 generally fell near 0.6, indicating that this is still a useful strategy for detecting CREs. However, the same paper found much stronger

correlations between the conventional MPRA design and other MPRA strategies, with Pearson's R^2 for average replicate activity across MPRA designs largely surpassing 0.8, indicating that the conventional orientation may be the more reliable for identifying reproducible CREs. (Fig. 2)

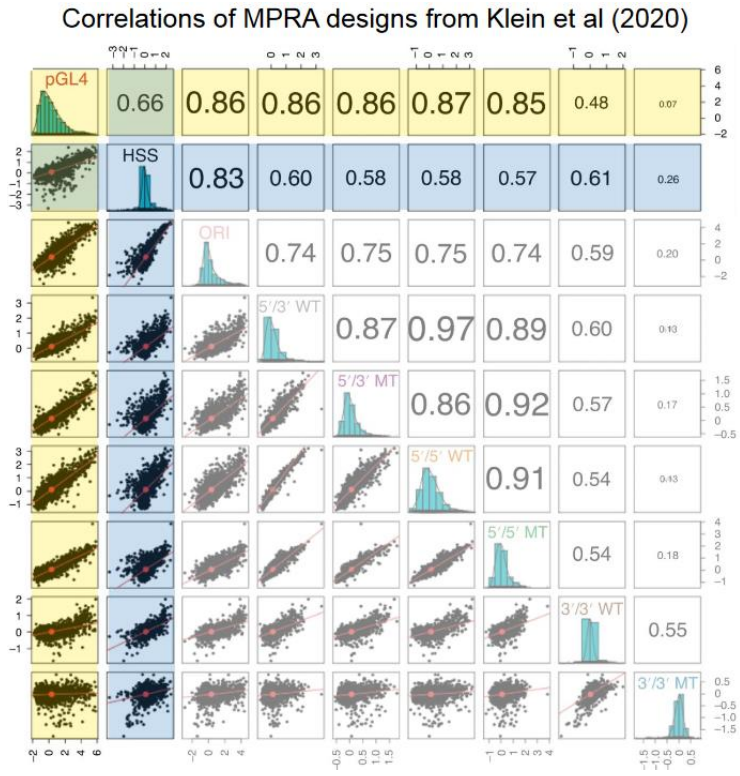


Fig. 2: Reproducibility of enhancers tested in different MPRA designs. Klein et al (2020)[11] tested nine different MPRA designs for a library of candidate enhancer elements. They averaged activity ratios for each enhancer across replicates within each design, then calculated Pearson's correlation coefficients for candidate enhancers' activity ratios across designs. Activity ratios obtained via STARR-seq ("HSS"), highlighted here in blue, showed moderately strong correlations with the other designs, with Pearson's R^2 generally near 0.6. Activity ratios obtained via the conventional design ("pGL4"), highlighted here in yellow, showed even stronger correlations, with Pearson's R^2 generally above 0.8. Other possible MPRA designs tested showed strong correlations, but are not discussed here as they were otherwise not suitable choices for the present work. (Figure adapted from Klein et al 2020[11])

In response to these concerns, I redesigned our lab's standard MPRA plasmid to use the conventional orientation and use barcodes which were carefully designed to minimize potential impacts on either RNA or PCR stability. I additionally cloned a complex library of barcodes into the same cloning site in the 3'UTR that was previously used for our STARR-seq experiments, and which is the technique employed in conventional-orientation MPRA plasmids. A standard approach for developing MPRA barcodes is to purchase synthetically produced oligonucleotides containing a string of randomized bases, such that all possible nucleotide combinations of a desired length may be represented as potential barcodes[12,13]. However, this can result in barcodes containing known RNA binding protein (RBP) sites, homopolymers, or variable GC content, which may impact RNA stability[14], PCR stability[15], or sequencing readout[16]. To minimize confounds in my barcode library while maximizing complexity and cloning modularity, I developed a strategy of creating barcode 'scaffolds' of predefined sequences punctuated by seven wildcard characters which, during synthesis, would be replaced with random nucleotides (e.g. *CG*ATC*GAC*TG*CTG*CT*CAA, where wildcards, denoted by asterisks, would be replaced with all possible combinations of nucleotides in different individual molecules). Starting with a small number of randomly designed scaffolds, I generated all possible sequences that would result when individual bases were inserted into the wildcard locations and when the barcodes were embedded into our MPRA plasmid sequence. I filtered scaffolds ones that would produce sites for known eukaryotic RBPs[17], nucleotide repeats greater than three bases in length, restriction enzyme sites relevant to our lab's cloning protocols, and GC content greater than 76% or less than 36% (respectively, 19 and 9 GC bases in the 25-base barcode, out of the total of 206 bp which would be amplified in my PCR product). I repeated this process until I had identified 25 scaffold sequences whose possible combinations would not violate any of these criteria when cloned into our MPRA plasmid. With 25 scaffold sequences, each containing seven wildcards that would be replaced with four nucleotides at random, this resulted in a library of up to 409,600 unique barcode sequences ($4^7 \times 5 = 409{,}600$). I cloned synthetic oligos

containing these scaffold sequences into my redesigned MPRA plasmid, generating several stocks of highly complex barcoded plasmid libraries for use for our lab's future MPRA experiments.

Using negative controls in MPRA experiments

Minimally active promoters are a core aspect of MPRA design, but many minimal promoters still experience some degree of transcriptional leakage[18,19]. As a result, negative controls are crucial for scientifically rigorous deployment of an MPRA as they enable estimation of background transcription levels in an experiment. This allows for confident identification of regulatory elements which are driving transcription to a degree significantly different from controls.

However, there is not yet a standard catalog of sequences which are verified to not show any regulatory activity, and publicly available epigenetic data from resources like ENCODE[20] have shown imperfect concordance with actual CRE activity as defined from functional testing[21]. This means our best guesses for MPRA control sequences are not guaranteed to actually serve as high-quality controls. Owing to the lack of clarity around negative control sequences, MPRA researchers sometimes choose to forego negative controls and instead only focus on the most active candidates to actually label and pursue as CREs[12]. While this approach has identified many enhancers that are supported by auxiliary evidence such as epigenetic modifications[12], it also almost certainly results in false positives and negatives. Further, this approach has historically resulted in disregarding the potential of MPRAs to also identify silencer sequences as, without an estimate of the basal transcription rate that occurs off of the MPRA plasmid by itself, it is extremely difficult to identify sequences that are producing *less* RNA than baseline. Fortunately, researchers are beginning to develop methods to identify silencer sequences via MPRA[22,23], as a growing body of literature is increasingly linking disrupted silencers to the development of psychiatric and other disease[24].

In order to improve the scientific rigor of MPRA experiments in our lab and aid in the unbiased identification of both activating and silencing regulatory sequences, I developed a library of negative controls to include in our MPRAs. I first tested these control sequences against the backbone MPRA construct into which I did not clone any CRE. The majority of my negative controls did not show a difference in transcriptional activity versus empty sequence in postnatal mouse brain. I describe the development, validation, and subsequent experimental use of this control library below.

References

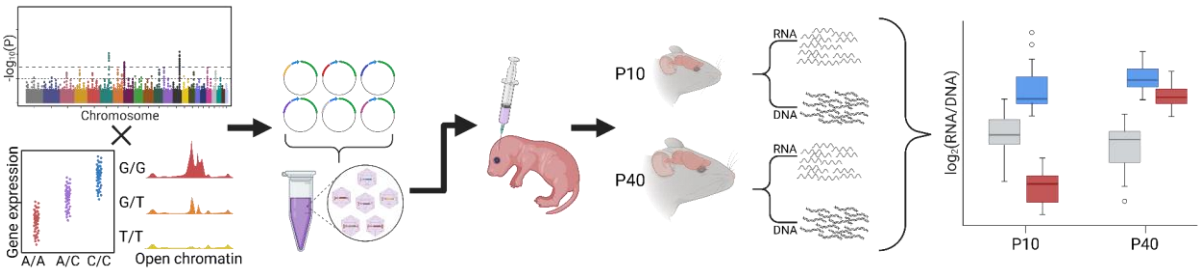
1. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology* (2012) doi:10.1038/nbt.2136.
2. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271–277 (2012).
3. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* (2013) doi:10.1126/science.1232542.
4. Nord, A. S. & West, A. E. Neurobiological functions of transcriptional enhancers. *Nature Neuroscience* (2020) doi:10.1038/s41593-019-0538-5.
5. Cheng, F. *et al.* Intronic enhancers of the human *SNCA* gene predominantly regulate its expression in brain *in vivo*. *Sci. Adv.* **8**, eabq6324 (2022).
6. Ott, C. J. *et al.* Intronic enhancers coordinate epithelial-specific looping of the active *CFTR* locus. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19934–19939 (2009).
7. Tuvikene, J. *et al.* Intronic enhancer region governs transcript-specific Bdnf expression in rodent neurons. *eLife* **10**, e65161 (2021).
8. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
9. Chan, Y.-C. *et al.* An unbiased AAV-STARR-seq screen revealing the enhancer activity map of genomic regions in the mouse brain *in vivo*. *Sci Rep* **13**, 6745 (2023).
10. Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol* **11**, a034728 (2019).
11. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**, 1083–1091 (2020).
12. McAfee, J. C. *et al.* Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. *Cell Genomics* **3**, 100404 (2023).
13. Zhao, S. *et al.* A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat Genet* **55**, 346–354 (2023).
14. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters* **582**, 1977–1986 (2008).
15. Mamedov, T. G. *et al.* A fundamental study of the PCR amplification of GC-rich DNA templates. *Computational Biology and Chemistry* **32**, 452–457 (2008).

16. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* **3**, lqab019 (2021).
17. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
18. Begeman, I. J., Emery, B., Kurth, A. & Kang, J. Regeneration and developmental enhancers are differentially compatible with minimal promoters. *Developmental Biology* **492**, 47–58 (2022).
19. Shapira, S. *et al.* Selective eradication of cancer cells by delivery of adenovirus-based toxins. *Oncotarget* **8**, 38581–38591 (2017).
20. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
21. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
22. Mouri, K. *et al.* Whole-genome functional characterization of RE1 silencers using a modified massively parallel reporter assay. *Cell Genomics* **3**, 100234 (2023).
23. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**, 1061 (2020).
24. Huang, D. & Ovcharenko, I. The contribution of silencer variants to human diseases. Preprint at https://doi.org/10.1101/2024.05.17.24307558 (2024).

c. Pre-publication manuscript

MPRA of schizophrenia-associated regulatory sequences in adolescent mouse brain

Graphical abstract



Created with BioRender.com

Abstract

Schizophrenia (SZ) risk variants identified from genome-wide association studies (GWASs) are highly enriched in synaptic genes and in presumed *cis*-regulatory elements (CREs) such as enhancers and silencers. A prominent hypothesis is that many of these risk variants disrupt the appropriate activity of their surrounding CREs. We conducted a massively parallel reporter assay (MPRA) in postnatal mouse brain to test this hypothesis for a set of prioritized SZ GWAS

risk variants. Variants were ascertained on strength of association to SZ, previous identification as chromatin accessibility quantitative trait loci (caQTLs) in human brain, and previous identification as expression quantitative trait loci (eQTLs) of synaptic genes in human brain. We injected our MPRA library into P0 mouse brain and collected DNA and RNA from cortex and hippocampus at P10 (juvenile mice) and P40 (adolescence). We compared CRE-driven activity to a set of empirically validated negative controls. 39 CREs showed regulatory activity at $p_{FDR} \leq 0.10$. 19 of these showed differences in activity between the reference and SZ risk allele. CREs with adolescent-specific upregulation were enriched for binding motifs for transcription factors (TFs) with postnatally dynamic expression. The binding motif for the SZ-associated TF Emx2 was enriched in CREs with allelic differences. We highlight three loci with allelic differences and adolescent-specific upregulation for detailed discussion. We additionally developed a set of validated negative controls for use in MPRAs and expanded the use of MPRA techniques to assess longitudinal postnatal changes in CRE activity. Our results provide evidence that SZ risk variants may disrupt CREs of particular relevance during adolescent neurodevelopment. These findings may help elucidate etiological pathways for SZ which may one day serve as treatment targets in adolescents at high risk for the disorder.

Introduction

Nearly twenty years of genome-wide association studies (GWASs) for schizophrenia (SZ) have made enormous gains into understanding the genetic components underlying SZ risk. Robust evidence has indicated that risk variants are enriched in loci which are thought to regulate gene expression[1–3]. Researchers have jumped to the task of characterizing these loci, using a range of methodologies to test the regulatory impact of SZ-associated common variants and the likely downstream biological impacts of aberrant gene regulation. Results have supported the hypothesis that many SZ risk alleles may alter the function of *cis*-regulatory elements (CREs)

such as enhancers[4,5] and silencers[6], and that aberrant expression of their targeted genes disrupts biological pathways that are known to be altered in SZ[7,8].

Given that the first symptoms of SZ typically become apparent during adolescence and early adulthood[9,10], SZ may present a promising opportunity among neurodevelopmental disorders to apply treatments or even preventative interventions postnatally. This may allow clinicians and researchers to avoid the oft-discussed methodological and ethical challenge of how to effectively apply preventative therapeutics during the prenatal period[11–14]. However, one of the most intriguing findings to come out of SZ genetic research is the apparent importance of SZ-associated CREs and genes during prenatal brain development[15,16]. The field faces a major challenge in reconciling this finding with the adolescent and adult timeframe for the onset of SZ symptoms.

A possible explanation for this discrepancy is that CREs impacted by SZ variants may regulate genes which are important to adolescent neurodevelopmental processes in addition to their role in prenatal development. A particularly critical aspect of adolescent neurodevelopment is synaptic pruning and remodeling[17]. Robust evidence indicates that SZ may specifically impact synapses[18–21]. As a consequence of these two observations, it has long been hypothesized that healthy adolescent synaptic pruning processes may go awry in SZ[22]. Given that CREs are responsible for dynamic changes in gene expression, including the regulation of genes in a manner that is specific to the developmental stage of their cell, tissue, or organism[23], it seems likely that they play a role in the appropriate execution of synaptic remodeling during adolescence. As a result, SNPs which disrupt synaptically relevant CREs during adolescence may increase the risk for synaptic disorders including SZ.

In the present study, we hypothesize that many noncoding SZ risk SNPs impact the appropriate activity of synaptically relevant CREs during adolescence. We examine CRE activity at a set of

prioritized noncoding SZ GWAS variants which have been previously identified as chromatin accessibility quantitative trait loci (caQTLs) and expression quantitative trait loci (eQTLs) of synaptic genes in postmortem human cortex. For these variants, we test the hypotheses that (1) they are located within CREs, (2) the risk allele at these variants alters the activity of the CRE, and (3) the CRE shows a temporally dynamic pattern of activity in postnatal mouse brain. Our results find that many SZ risk variants are indeed located within elements that show *cis*-regulatory activity in postnatal mouse brain. Subsets of active CREs show differential activity in the presence of the SZ risk allele and specific trajectories of activity from preadolescent to adolescent mouse brain development. TF motif analyses provide insights as to possible mechanisms of postnatally dynamic activity and differences between reference and risk haplotypes. We highlight three loci which show allelic differences in gene regulation as well as specific induction of activity during adolescence. We additionally intersect our findings with publicly available datasets to generate hypotheses on likely gene targets of these loci. These findings shed light on the role that noncoding SZ risk variants may have in disrupting critical brain processes, including during postnatal brain development, and may help elucidate postnatally actionable therapeutic targets for children and teenagers at risk for developing SZ.

Results

Negative control library

To control for background transcription rates off the MPRA plasmid, we designed a library of 25 negative control candidates to include in the present experiments. Of these candidates, 15 were human-derived elements in the third intron of the gene *CACNA1C* that did not show regulatory activity in our previous MPRA in postnatal mouse brain[24]. Ten additional intergenic elements were selected at random from mm10 that showed no evidence *in silico* of regulatory function and were cloned out of wild-type C57/BL6J mouse DNA. These 25 elements were cloned into MPRA plasmids containing the β-globin minimal promoter[25] and a library of barcode elements in

the 3' UTR (Fig. 1A). Barcodes were designed to minimize potential MPRA confounds, including eukaryotic RNA binding protein motifs[26], single-nucleotide repeats greater than three bases in length, and GC content. Elements were matched to a median of 261 and a mean of 310 barcodes.

This library was co-packaged into an AAV of serotype PHP.eB along with a small amount of MPRA plasmid into which no regulatory element was cloned ("empty plasmid", represented by 97 barcodes) (Fig. 1A). Mice were injected with the control virus on postnatal day 0 (P0) and bulk cortex and hippocampus were collected on P28. RNA and DNA were sequenced from each sample and barcodes identified in the sequencing reads were matched back to their associated regulatory elements. Activity ratios were calculated for each control candidate as well as for the empty plasmid. Ratios were regressed against sex and the number of barcodes detected for each CRE. Regulatory activity for each candidate was modeled against that of empty plasmid. Of the 25 tested candidates, 24 were determined to be detected sufficiently for testing. Two showed significantly reduced activity versus empty plasmid at an unadjusted $p \leq 0.05$ (Fig. 1B). While these did not pass multiple testing corrections, these regulatory elements were not retained for subsequent MPRA analyses in order to ensure that only elements with strong evidence for neutral regulatory activity were used as controls. The remaining 22 elements, as well as empty plasmid, were retained in all subsequent analyses as empirically validated negative controls.

SZ variant library cloning

Summary statistics from the most recent SZ GWAS from the Psychiatric Genomics Consortium (PGC)[18] were reviewed and variants were prioritized based on strength of association with SZ. Variants were subsetted to those that had been previously identified as caQTLs[27] and eQTLs[28] of synaptic genes in postmortem human brain. After dataset intersection, 152 common variants were prioritized across 139 putative CREs of approximately 300-500 bp in length. Six CREs,

containing seven variants, were discarded due to complex repeat regions that complicated primer design. Three variants, within three CREs, were excluded due to high proximity to exon regions. Of the remaining 130 loci, 93 CREs containing 100 variants were successfully cloned into the MPRA library, and 78 of these CREs showed successful cloning of both the reference and risk alleles (Fig. 1C). CREs were amplified out of genotyped DNA from healthy humans[29] who were homozygous for either the reference or SZ risk haplotypes across linkage disequilibrium (LD) blocks containing the putative CREs. They were then cloned into our barcoded MPRA plasmid library. With alleles considered individually, each CRE was matched to a median of 163 and a mean of 297 barcodes. CREs were represented relatively evenly in overall library composition (Fig. 1D), ensuring strong capability for candidate testing in postnatal mouse brain.

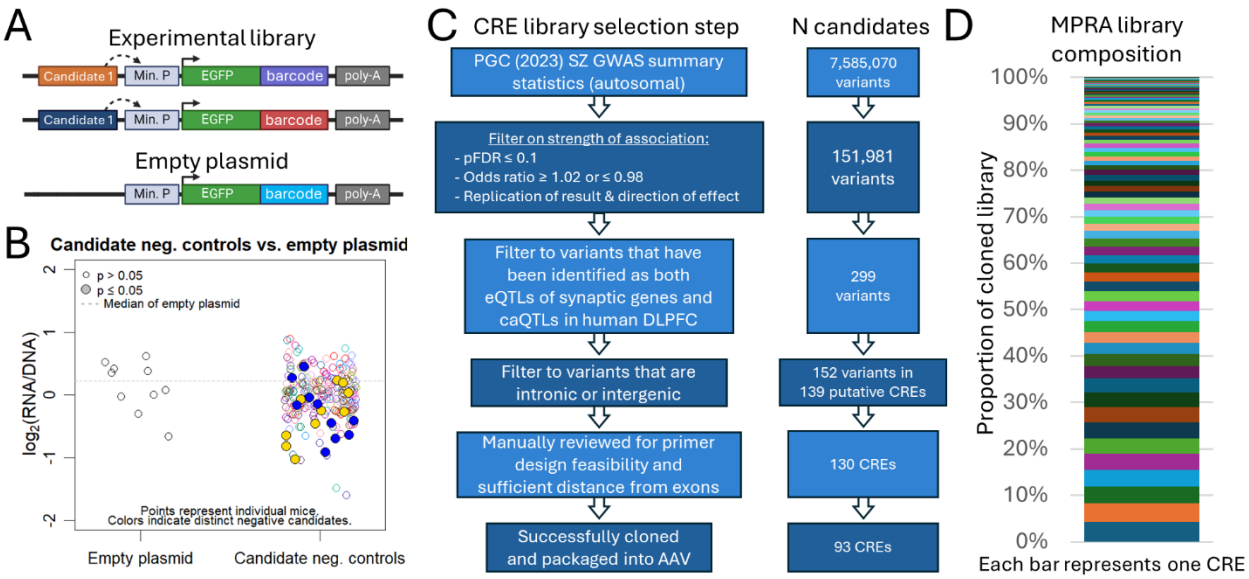


Fig. 1: Experimental setup. (A) Our CRE-reporter construct consisted of the fluorescent reporter gene EGFP under the control of the β-globin minimal promoter[25]. CRE candidates were cloned upstream of the promoter and 25 bp barcodes were cloned into the 3' UTR of EGFP. For negative control testing, an additional construct was included into which we did not clone any candidate CRE element ("empty plasmid"). (B)

Negative controls were tested for regulatory activity in postnatal mouse brain versus empty plasmid. Two of the 24 candidate negative controls showed significantly reduced activity versus empty plasmid at unadjusted $p \leq 0.05$, though these did not pass multiple testing corrections. The remaining 22 candidates as well as empty plasmid were retained as validated negative controls for all downstream experiments. (C) We filtered the 7.5 million variants from the most recent PGC SZ GWAS autosomal summary statistics (Trubetskoy et al, 2022)[18] on strength of association as well as evidence for affecting CREs of synaptic genes in human DLPFC. 93 candidate CREs containing SZ risk SNPs were eventually cloned into the final library. (D) CREs were relatively evenly represented in the final cloned MPRA library. (Min. P: minimal promoter)

SZ-associated CREs show activity in preadolescent and adolescent mouse brain

To assess detectability of CRE-driven activity off of the MPRA plasmid in both preadolescent and adolescent brain, mice were injected at P0 with the SZ MPRA library, which was packaged into an AAV of serotype PHP.eB. They were also co-injected with a constitutively expressed CAG-mRuby3 positive control packaged into PHP.eB-AAV. Brains were collected at P10 and P40. Both EGFP and mRuby3 constructs showed activity at both timepoints (Fig. 2A,B), indicating that CRE-driven expression was detectable in our experiments at P10 and P40.

For sequencing experiments, the SZ MPRA library was co-injected with the negative control library, including empty plasmid, at P0. Bulk cortex and hippocampus were collected at P10 and P40 for RNA and DNA sequencing. Barcodes in the sequencing reads were matched back to their associated regulatory elements, and activity ratios were calculated for each element. Ratios were regressed against sex and number of barcodes detected per CRE. RNA reads per CRE and DNA reads per CRE showed strong correlations across mice at each collection timepoint (Fig. S1A,B). Activity ratios across mice, within-age, showed more moderate

correlations (Fig. S1C), consistent with correlations reported in previous MPRAs[30]. We next tested for concordance of elements with strong evidence of regulatory activity, represented by a doubling or halving of activity relative to negative controls ($\log_2(\frac{RNA}{DNA})$ ±1). For a given CRE which met this threshold, we calculated the proportion of mice which showed the same direction of CRE activity versus negative controls (Fig. 2C). For example, if 8 out of 10 mice showed strong upregulation from a CRE, and 2 out of 10 mice showed strong downregulation, that CRE was determined to be 80% concordant. CREs showed a high degree of concordance across mice relative to a null distribution (P10, n CREs = 28; P40, n CREs = 42) (One-tailed Wilcoxon sign rank test, P10: p =4.22e-09; P40: p = 1.90e-10) (Fig. 2D,E). This suggests that performance improves when detecting either strong enhancers or strong silencers.

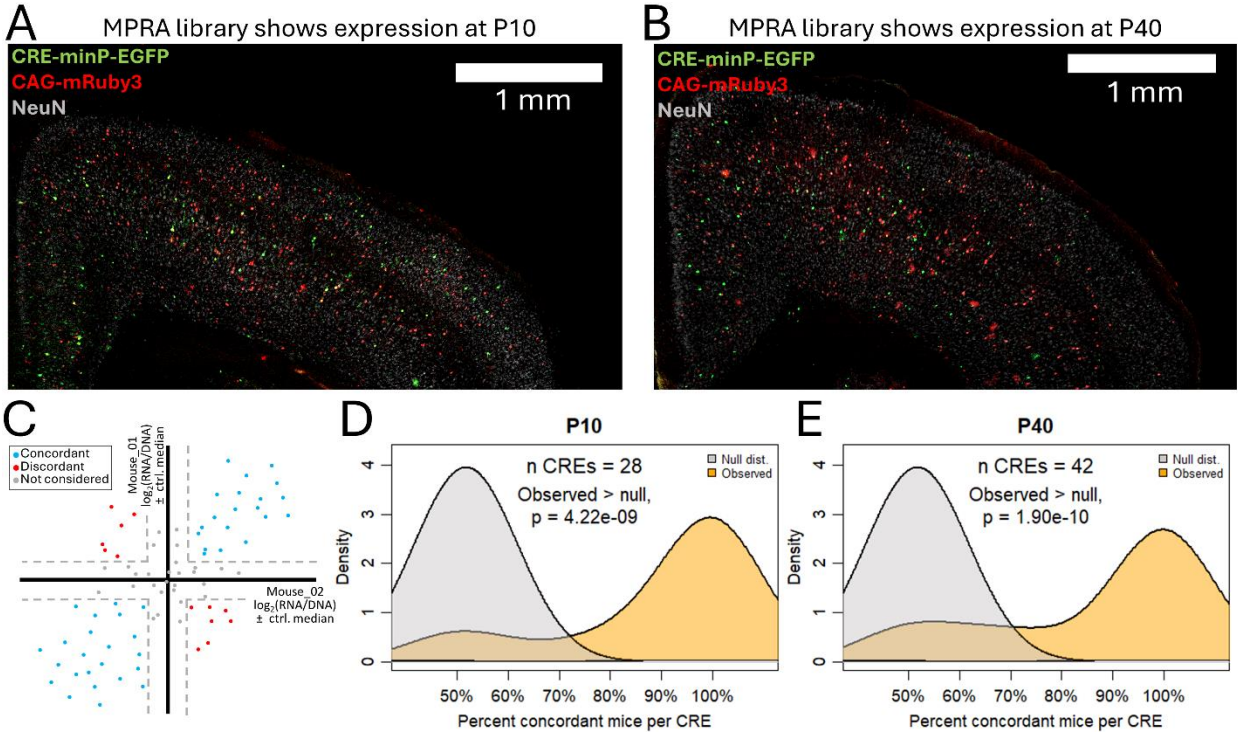


Fig. 2: CREs show reproducible activity in postnatal mouse brain. (A,B) The library of CREs drives EGFP expression in mouse cortex at (A) P10 and (B) P40. The positive control CAG-mRuby3 AAV also drives expression at both ages. (C) Schematic showing methodology for concordance calculations. For each mouse, CREs were first subset into

those showing strong enhancer or silencer activity by removing those with $\log_2(\frac{RNA}{DNA})$ within ±1 of the median of the negative controls. For each remaining CRE, mice showing activity in the same direction were considered concordant, while mice showing activity in opposite directions were considered discordant. The percent of concordant mice was calculated for each CRE. (D, E) CREs with strong activity showed substantial concordance at both P10 (D) and P40 (E), while a null distribution performed at chance.

Regulatory candidates were first tested for allelic differences in activity. 48 CREs showed sufficient detection of both alleles to test for differences. Allelic differences were modeled both controlling for age at collection and as an interaction effect with age. Given that SNPs were ascertained on previous identification as both eQTLs and caQTLs, we anticipated high rates of allelic differences in our assay. Indeed, 25 CREs (52%) showed allelic differences at $p \leq 0.05$. After multiple testing corrections, 21 (44%) remained significant at an FDR-adjusted p-value (p_{FDR}) ≤ 0.10 and were retained as separate alleles for all subsequent analyses (Fig. 3A). For all other elements, barcodes representing separate alleles were combined and the CRE was considered as a single candidate for all subsequent analyses.

CREs were next modeled for activity against negative controls, both controlling for and as an interaction with age at collection. 73 CREs were determined to have sufficient detection across both ages in order to test. 45 CREs (62%) showed a nominally significant difference in regulatory activity versus negative controls, and 39 (53%) of these passed multiple testing corrections at $p_{FDR} \leq 0.10$. These CREs included both enhancer and silencer elements, several of which showed dynamic regulatory trajectories across timepoints (Fig. 3A). We plotted activity distributions for enhancers, silencers, and negative controls. Distributions were unimodal and normally distributed around their peaks. Enhancers showed a clear shift to the right versus negative controls, indicating increased activity, while silencers were shifted to the left, indicating

reduced activity (Fig. 3B). We report all $p_{FDR} \leq 0.10$ results in Table 1 and Fig. S2, and all additional results in Supplementary Table 1.

We asked whether CREs which showed greater transcriptional activity during adolescence than in juvenile mice might be enriched for binding sites for postnatally dynamic TFs. We conducted a TF motif enrichment analysis of significantly active CREs which showed upregulation during adolescence versus sequences which were neutral (unadjusted $p > 0.05$) in our assay. Indeed, CREs with upregulation during adolescence were enriched for binding sites for Jund (5/11 CREs, p = 1.86e-03), Nfix (4/11 CREs, p = 6.53e-03), and Znf266 (7/11 CREs, p = 6.53e-03) (Fig. 3C). Both *Nfix* and *Zfp266* (mouse ortholog to *ZNF266*) undergo strong downregulation from juvenile to adolescent mouse brain (Fig. 3D). Zfp266 is predominantly involved in transcriptional repression[31], while Nfix acts as both an activator and a repressor of its targeted genes[32]. Reduced expression of these TFs into mouse adolescence may result in reduced repression on those CREs which preferentially bind them; this phenomenon may mediate some of the upregulation observed at adolescent-specific enhancers in our assay.

We then asked whether active CREs which showed allelic differences might show enrichment of binding sites for SZ-associated TFs. Of the 21 CREs with FDR-significant allelic differences, 19 of these also showed FDR-significant activity in the overall assay versus negative controls. We conducted a TF enrichment analysis for these active CREs with allelic differences (n = 19) versus sequences which were neutral for activity (n = 28). CREs which featured allelic changes were significantly enriched for binding motifs for the TFs Emx2, Gsx2, Myod1, and Znf667 (all TFs, 6/19 CREs, p = 5.49e-03) (Fig. 3F). Perhaps most interesting among enriched TFs is Emx2. This TF is critical for the developing CNS[33] and is highly expressed in cortex from embryonic through adult neurodevelopment in mice[34] as well as in human adult cortex[35]. *EMX2* is overexpressed in DLPFC from humans with SZ[36] and has been identified as a potential hub gene which may regulate a network of co-dysregulated genes in SZ[37]. It is possible that the

biological pathways which link *EMX2* and a subset of our CREs to SZ development may be at least partially overlapping. This hypothesis may merit future research in order to better define the processes which may mediate the relationship between SZ and *EMX2* as well as our identified allelic CREs.

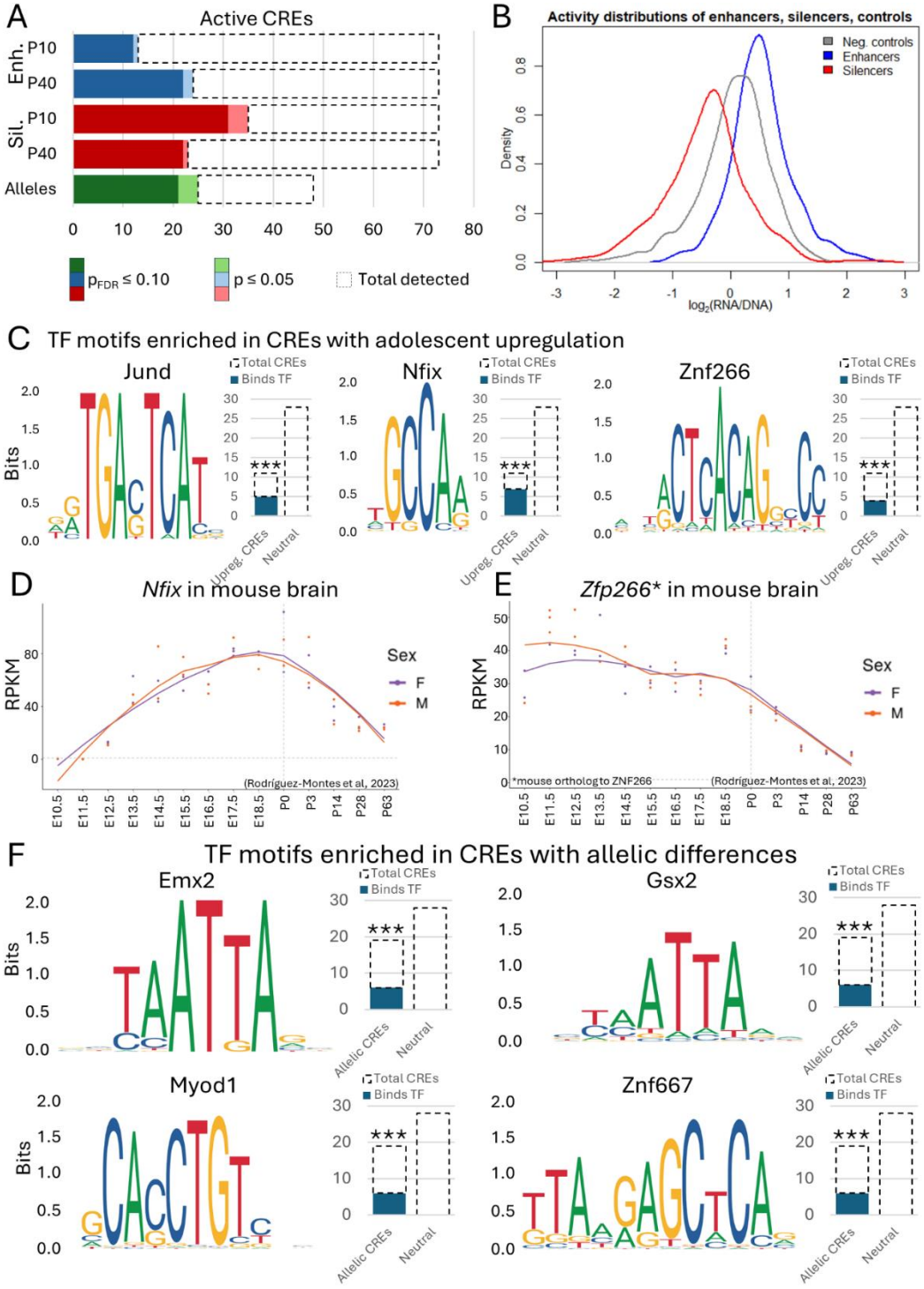


Fig. 3: Summary of activity in CRE library. (A) Of 73 CREs which were detected sufficiently for testing, a large portion showed either enhancer or silencer activity in our library at $p \leq 0.05$ and $p_{FDR} \leq 0.10$. Of the 48 CREs with sufficient detection of both alleles, approximately half showed differences in activity between the risk and reference alleles. (B) FDR-significant enhancers and silencers, as well as negative controls, showed normal distributions for activity. Enhancers were shifted higher in activity than controls, while silencers were shifted toward reduced activity. (C) Binding sites for the TFs Jund, Nfix, and Znf266 were enriched in CREs with adolescent-specific upregulation versus candidate CREs which were neutral for activity in our assay. Out of the eleven adolescent-upregulated sequences, five contained binding motifs for Jund (45%), four contained motifs for Nfix (36%), and seven contained motifs for Znf266 (64%). None of these motifs were present in the 28 neutral sequences. (D) Nfix, a bidirectional TF, undergoes strong downregulation from preadolescent to adolescent and adult mouse brain, (E) as does Zfp266 (mouse ortholog to Znf266), which has repressive activity. (Plots generated using the Evo-Devo Sex-Biased Mammalian Genes app[38]) (F) Binding sites for the TFs Emx2, Gsx2, Myod1, and Znf667 were enriched in active CREs which showed allelic differences. Six of 19 (32%) active allelic CREs contained binding motifs for each TF, while these motifs were present in zero out of the 28 neutral sequences. (Sil.: silencer, Enh.: enhancer)

Table 1: Results for all CREs active at $p_{FDR} \leq 0.10$

CREs	Coordinates (hg38)	SNP(s)	Risk allele(s)	Ref. allele(s)	Alleles tested?	pFDR for allelic differences		CRE activity											
								Without alleles				Risk allele				Reference allele			
								Controlling for age		Interaction with age		Controlling for age		Interaction with age		Controlling for age		Interaction with age	
						Ctrl. for age	Interaction with age	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p
CRE_002	chr1: 27148853-27149295	rs11247612	C	T	FALSE	NA	NA	0.55	0.00	0.02	0.10	NA	NA	NA	NA	NA	NA	NA	NA

CRE_005	chr1: 205079950-205080424	rs9651228	G	A	FALSE	NA	NA	0.15	0.45	0.04	0.00	NA	NA	NA	NA	NA	NA	NA	NA
CRE_014	chr2: 48337966-48338436	rs59761994	G	A	TRUE	0.07	0.10	NA	NA	NA	NA	-0.54	0.00	-0.03	0.02	0.12	0.52	0.01	0.39
CRE_015	chr2: 48339556-48340049	rs6545034	C	T	TRUE	0.20	0.53	-0.32	0.05	0.02	0.13	NA	NA	NA	NA	NA	NA	NA	NA
CRE_016	chr2: 48341409-48341886	rs62137002	G	A	TRUE	0.00	0.75	NA	NA	NA	NA	0.05	0.80	0.01	0.33	0.94	0.00	0.02	0.13
CRE_017	chr2: 48354574-48355005	rs11685861	T	C	TRUE	0.08	0.76	NA	NA	NA	NA	-0.83	0.00	-0.02	0.12	0.42	0.03	0.00	0.99
CRE_020	chr2: 48358463-48358924	rs7563625	A	C	FALSE	NA	NA	0.45	0.00	0.02	0.08	NA	NA	NA	NA	NA	NA	NA	NA
CRE_030	chr4: 87523885-87524351	rs4693832	A	G	TRUE	0.01	0.90	NA	NA	NA	NA	-0.39	0.01	0.01	0.50	0.24	0.16	0.01	0.39
CRE_034	chr5: 64401976-64402366	rs898519	G	A	TRUE	0.00	0.26	NA	NA	NA	NA	0.12	0.50	0.01	0.56	-0.30	0.07	0.00	0.89
CRE_035	chr5: 153924782-153925243	rs13170510	G	A	TRUE	0.81	0.26	-0.48	0.00	0.00	0.83	NA	NA	NA	NA	NA	NA	NA	NA
CRE_038	chr5: 153951427-153951885	rs35428024	G	A	TRUE	0.04	0.58	NA	NA	NA	NA	-0.52	0.00	-0.01	0.71	0.20	0.28	0.01	0.40
CRE_040	chr5: 153963255-153963754	rs58798017, rs66607728	T, C	C, G	FALSE	NA	NA	-0.30	0.08	0.02	0.19	NA	NA	NA	NA	NA	NA	NA	NA
CRE_042	chr5: 153987019-153987412	rs425263	A	G	FALSE	NA	NA	0.37	0.02	0.01	0.35	NA	NA	NA	NA	NA	NA	NA	NA
CRE_043	chr5: 153988354-153988751	rs453034	G	T	FALSE	NA	NA	-0.21	0.23	0.03	0.00	NA	NA	NA	NA	NA	NA	NA	NA
CRE_045	chr5: 154009899-154010396	rs12651907	A	G	TRUE	0.50	0.26	-0.01	0.97	0.02	0.06	NA	NA	NA	NA	NA	NA	NA	NA
CRE_051	chr6: 31040643-31041102	rs2517549	A	C	TRUE	0.01	0.81	NA	NA	NA	NA	0.51	0.00	0.01	0.46	-0.64	0.00	0.00	0.97
CRE_052	chr6: 31041230-31041704	rs2523901	C	T	FALSE	NA	NA	-0.56	0.00	-0.02	0.14	NA	NA	NA	NA	NA	NA	NA	NA
CRE_055	chr6: 31043055-31043554	rs2844647, rs2248386	A, T	G, C	TRUE	0.10	0.48	NA	NA	NA	NA	0.38	0.02	0.00	0.97	-0.25	0.19	-0.02	0.16
CRE_062	chr6: 31049481-31049854	rs2523866, rs2844643	C, A	T, G	FALSE	NA	NA	0.34	0.06	0.03	0.01	NA	NA	NA	NA	NA	NA	NA	NA
CRE_063	chr6: 31050404-31050890	rs2523865, rs2523864	C, T	A, C	FALSE	NA	NA	-0.44	0.06	0.00	0.83	NA	NA	NA	NA	NA	NA	NA	NA
CRE_069	chr6: 87898106-87898580	rs9450822	G	A	TRUE	0.04	0.19	NA	NA	NA	NA	0.52	0.00	0.01	0.37	0.06	0.75	0.03	0.00
CRE_070	chr6: 87900187-87900655	rs4235834	G	T	TRUE	0.00	0.05	NA	NA	NA	NA	0.24	0.16	0.02	0.06	-0.36	0.02	0.00	0.97
CRE_073	chr6: 100605037-100605531	rs240115	C	T	TRUE	0.78	0.23	-0.32	0.06	0.01	0.40	NA	NA	NA	NA	NA	NA	NA	NA
CRE_079	chr6: 100632627-100633112	rs240143	A	G	TRUE	0.21	0.84	-0.20	0.27	0.03	0.02	NA	NA	NA	NA	NA	NA	NA	NA
CRE_082	chr6: 100640959-100641447	rs239246	A	G	TRUE	0.00	0.23	NA	NA	NA	NA	0.17	0.33	0.02	0.16	-0.79	0.00	0.00	0.98
CRE_083	chr6: 100642200-100642573	rs239245	A	C	TRUE	0.06	0.54	NA	NA	NA	NA	-0.13	0.48	0.01	0.30	-0.49	0.00	0.00	0.76
CRE_093	chr7: 127475956-127476260	rs6957647	A	G	FALSE	NA	NA	0.36	0.07	0.01	0.45	NA	NA	NA	NA	NA	NA	NA	NA

CRE_099	chr7: 127493877-127494356	rs34976296	T	C	TRUE	0.72	0.03	NA	NA	NA	NA	-0.48	0.02	-0.04	0.01	-0.69	0.00	0.02	0.05
CRE_100	chr7: 127494107-127494580	rs13228307	A	G	TRUE	0.01	0.78	NA	NA	NA	NA	0.07	0.74	0.00	0.86	-1.01	0.00	-0.01	0.40
CRE_101	chr7: 127498321-127498799	rs1419401	G	T	FALSE	NA	NA	-0.02	0.95	0.03	0.01	NA	NA	NA	NA	NA	NA	NA	NA
CRE_103	chr7: 127501956-127502455	rs6975687, rs6979978	A, C	G, G	TRUE	0.72	0.05	NA	NA	NA	NA	-0.57	0.00	-0.01	0.26	-0.63	0.00	0.01	0.54
CRE_107	chr8: 22411059-22411542	rs11780055	A	T	TRUE	0.04	0.74	NA	NA	NA	NA	0.12	0.52	0.01	0.42	-0.58	0.01	0.00	0.97
CRE_110	chr8: 22431845-22432148	rs4872497	C	G	TRUE	0.88	0.85	0.19	0.29	0.03	0.02	NA	NA	NA	NA	NA	NA	NA	NA
CRE_112	chr8: 27400371-27400863	rs2163174	A	G	FALSE	NA	NA	-0.74	0.00	0.00	0.80	NA	NA	NA	NA	NA	NA	NA	NA
CRE_121	chr10: 48831865-48832352	rs7908454	G	A	TRUE	0.01	0.90	NA	NA	NA	NA	-0.51	0.00	0.01	0.50	0.23	0.19	0.01	0.68
CRE_126	chr17: 21379811-21380270	rs1317349	T	G	TRUE	0.63	0.37	0.48	0.00	0.01	0.46	NA	NA	NA	NA	NA	NA	NA	NA
CRE_127	chr17: 39276765-39277245	rs59585619	G	C	TRUE	0.02	0.90	NA	NA	NA	NA	-0.35	0.08	0.01	0.33	0.43	0.01	0.01	0.38
CRE_128	chr17: 39277370-39277868	rs56058902	T	C	TRUE	0.26	0.86	-0.24	0.20	0.02	0.08	NA	NA	NA	NA	NA	NA	NA	NA
CRE_130	chr18: 53429778-53430268	rs2879527	C	T	TRUE	0.06	0.75	NA	NA	NA	NA	-0.09	0.65	0.02	0.08	0.37	0.02	0.01	0.28

Potential gene targets of CREs altered by SZ alleles

Following general analysis of activity in our CRE library, we sought to generate hypotheses on possible targeted genes of CREs which showed allelic differences in activity in our assay. Of the 39 FDR-significant CREs, 19 showed FDR-significant allelic differences. We highlight here three loci which show evidence for both adolescent-specific upregulation of their targeted genes, as well as strong differences in activity between the risk and reference alleles. We intersected our results at these regulatory regions with publicly available epigenetic and gene expression datasets. Based on identified correlations, we generate preliminary hypotheses on gene regulatory mechanisms which may be relevant to the association between tested CREs, SZ, and adolescent neurodevelopment. While these relationships are not directly tested in the present study, they may merit testing in future work.

The first regulatory locus we highlight for detailed analysis spans six active regulatory elements in close proximity with each other on chromosome 5: CREs #35, 38, 40, 42, 43, and 45 (Fig. 4A). All CREs were represented by a range of barcodes and no barcode appeared to dominate the activity for any given CRE (Fig. S3A). Of CREs in this locus, only CRE #38 was identified as showing allelic differences in our assay. Our results found that the risk allele at CRE #38 induced transcriptional silencing, while the reference allele was apparently neutral for activity. The additional CREs in this locus which were not identified as showing allelic differences showed both enhancer and silencer activity. CREs #43 and 45 specifically showed upregulation of expression during adolescence, such that the cumulative regulatory effect of CREs across this locus changed from preadolescent silencing to adolescent enhancer activity in the reference haplotype. In the risk haplotype, this effect was shifted downward due to the silencing activity of the risk allele at CRE #38, such that the risk haplotype induced strong silencing in preadolescent mice, which was somewhat alleviated during adolescence (Fig. 4B). TF motif analysis shows that the SZ risk SNP at CRE #38, rs35428024 (reference: G, risk: A), alters the binding site of a number of TFs which are expressed in both mouse and human postnatal cortex. Specifically, the risk allele creates binding sites for Znf143, Neurod2, Foxp3, Ndf1, Ptf1a, and Tyy1, which collectively confer a range of activating and repressing functions on their targeted genes[39–44]. These binding sites are broken by the reference allele (Fig. 4C). The SNP in #38 is an eQTL in which the risk allele is associated with reduced expression of nearby genes *GRIA1* and *FAM114A2*[28], consistent with our identification of silencing activity from the risk allele. The CREs at this locus also show contacts with the *GRIA1* promoter in human neurons and microglia[45] (Fig. 4D). Both *GRIA1* and *FAM114A2* show evidence of upregulation into adolescence and adulthood in human postmortem brain[46] (Fig. 4E). This locus also shows chromatin contacts with the promoter for nearby *GALNT10*[45] (Fig. 4D). Based on these intersections with external datasets, we identify correlational evidence for *GRIA1*, *FAM114A2*, and *GALNT10* as possible gene targets of these CREs. These relationships should be explored

in future research to better map the biological pathways that may link the SZ SNP at CRE #38 to symptom development.

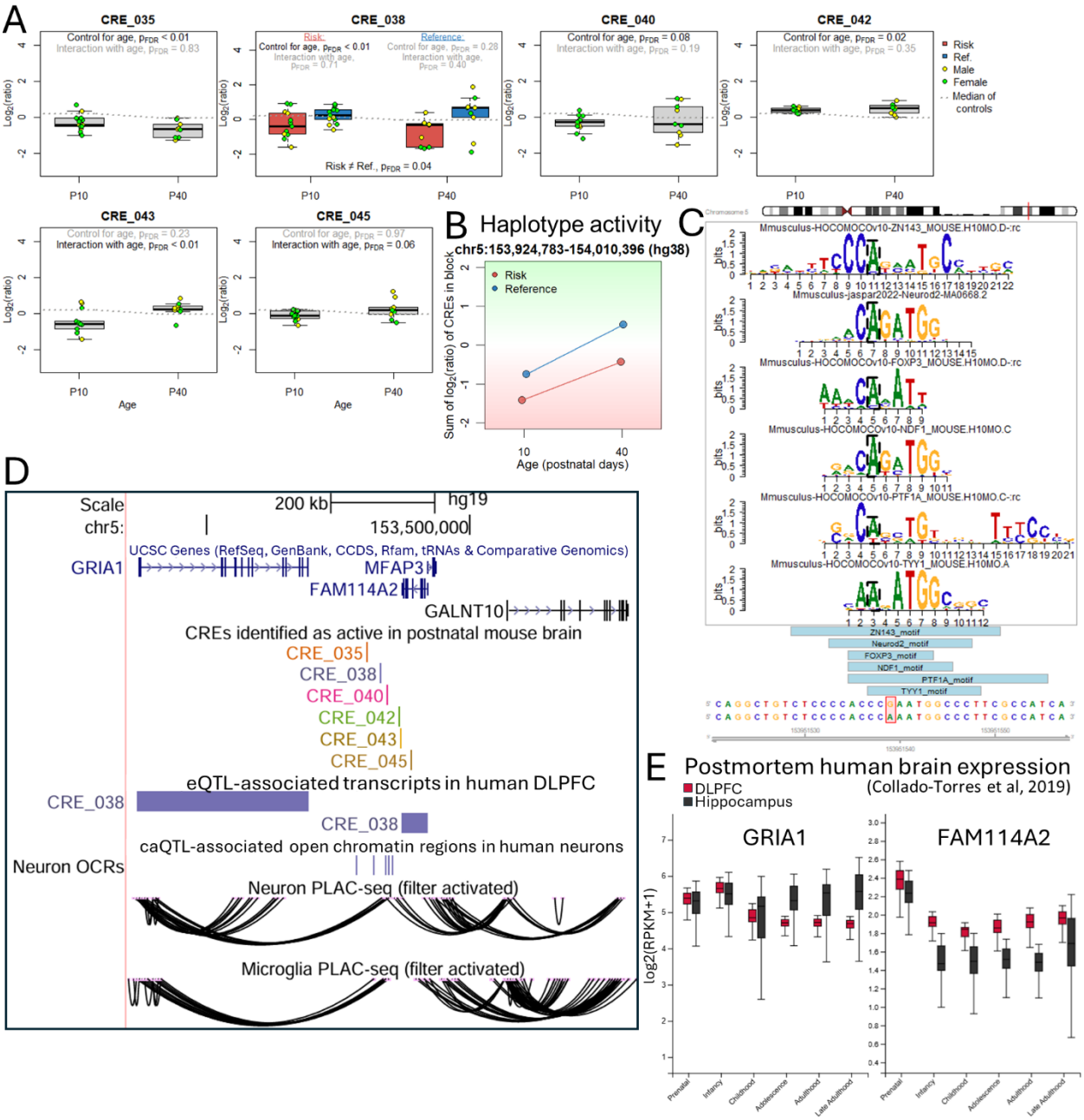


Fig. 4: Regulatory block on chromosome 5 shows adolescent upregulation and allelic differences. (A) Five CREs in high proximity to each other on chromosome 5 showed activity in our assay. Of these, only CRE #38 showed allelic differences. The

risk allele showed silencing at both ages studied, while the reference allele was neutral for activity. CREs #35 and 40 acted as silencers, CRE #42 acted as an enhancer, and CREs #43 and 45 showed upregulation between P10 and P40 mice. (B) The cumulative regulatory activity across this block showed upregulation from preadolescent to adolescent mice. The reference haplotype was a strong silencer in P10 mice and relatively neutral in P40 mice, while the risk haplotype was a silencer in P10 mice and an enhancer in P40 mice. (C) The risk SNP at CRE #38 (rs35428024, reference: G, risk: A) alters the binding site of a number of TFs. Specifically, several binding motifs are created by the risk allele which are not present with the reference allele. (D) The SNP in CRE #38 is a caQTL associated with local neuronal chromatin accessibility[27], as well as an eQTL in which the risk allele is associated with reduced transcription of *GRIA1* and *FAM114A2*[28]. The CREs in this locus show chromatin contacts in human neurons and microglia with the promoters for *GRIA1* and *GALNT10*[45]. (E) *GRIA1* and *FAM114A2* undergo slight upregulation in human adolescent and adult brain, consistent with the adolescent-specific upregulation at this regulatory locus in our assay (boxplots from the BrainSeq Phase 2 Developmental Expression Browser[46]).

Another locus of interest in our assay includes two CREs on chromosome 17: CREs #127 and 128. Both CREs were represented by a range of barcodes and no barcode appeared to dominate the activity for either CRE (Fig. S3B). At CRE #127, the risk allele induced transcriptional silencing at both ages tested, while the reference allele induced transcriptional activation. CRE #128 did not show allelic differences, but did show upregulation from P10 to P40 (Fig. 5A). TF motif analysis shows that the reference allele at CRE #127 (SNP: rs5958561, reference: G, risk: C) creates a binding site for the transcriptional activator[47,48] Prdm15 which is broken by the risk allele (Fig. 5B). The cumulative effect at the reference haplotype was relatively neutral activity during preadolescence and enhancer activity during adolescence, while

the risk haplotype shifted this trajectory downward, showing repression during preadolescence and relatively neutral activity during adolescence (Fig. 5C). The SNP in CRE #127 is an eQTL in which the risk allele is associated with downregulation of nearby genes *ARHGAP23* and *CDK12*, which is consistent with the direction of effect observed in our assay, but upregulation of *FBXL20* and *PPP1R1B*[28]. Chromatin conformation data shows contacts between these CREs and the putative promoters for *ARHGAP23*, *FBXL20*, and *CDK12*[45,49] (Fig. 5D). Of possible target genes, *CDK12* shows upregulation into adulthood in postmortem human brain[46] (Fig. 5E). It is possible that eQTL associations between the SZ risk allele and upregulation of *PPP1R1B* and *FBXL20* may be caused by other nearby eQTLs which are in linkage disequilibrium with the eQTL tested in CRE #127, or by biological contexts not replicated in our assay. More work is necessary to fully characterize the entire regulatory landscape of this SZ risk-associated region. However, given the intersections of chromatin contact data, eQTL data, TF motif analysis, and our finding of silencing from the risk haplotype, *ARHGAP23* and *CDK12* may show sufficient correlational evidence as possible gene targets of CREs #127 and 128 to merit direct testing of these relationships.

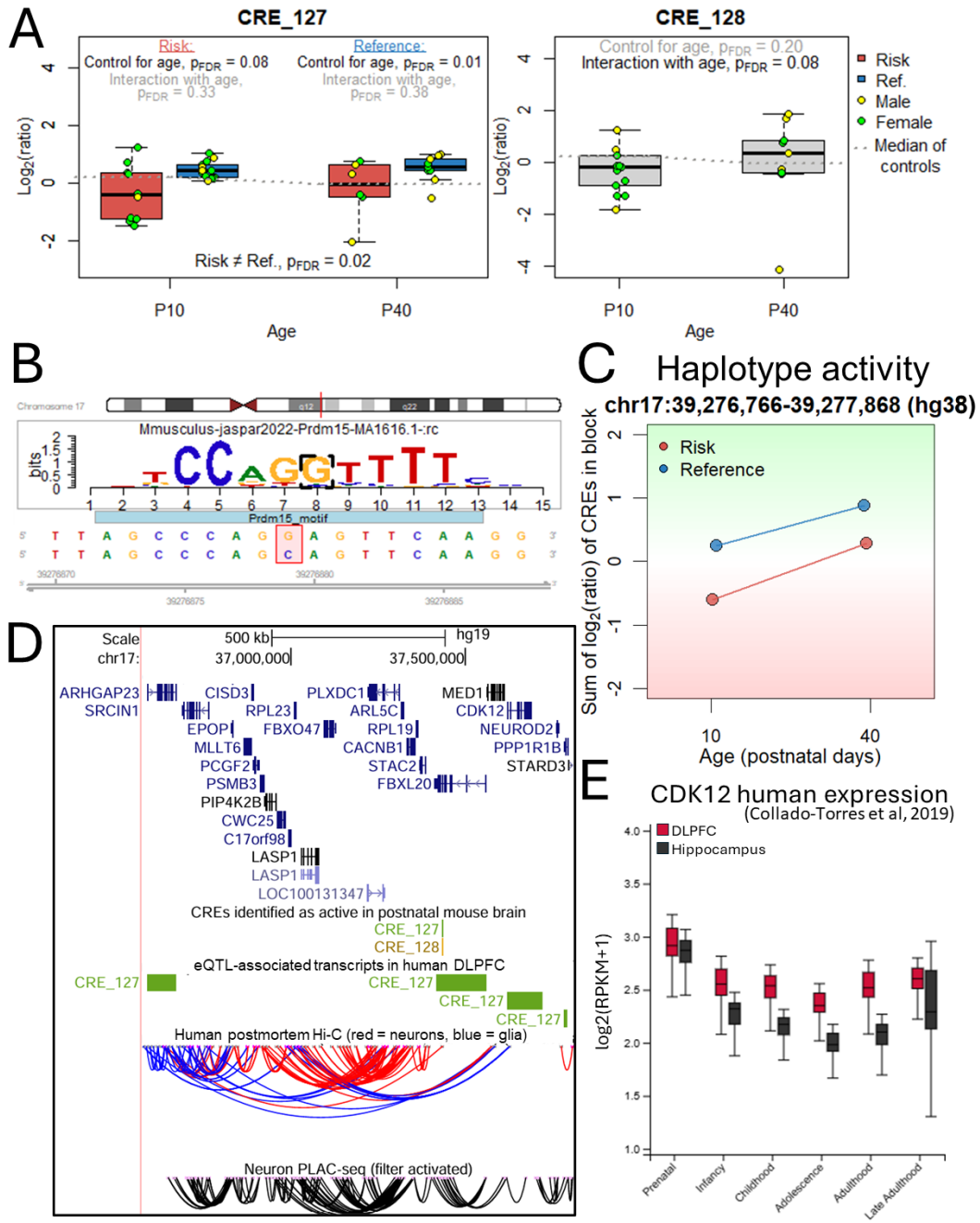


Fig. 5: Two adjacent CREs on chromosome 17 may exert local and long-ranging *cis*-regulatory influences. (A) CREs #127 and 128 showed activity in our assay. The reference allele at CRE #127 induced enhancer activity at both ages, while the risk allele induced silencer activity at both ages. CRE #128 did not show allelic differences, but did show upregulation from P10 to P40 mice. (B) The risk allele at the SNP in CRE #127 (rs5958561, reference: G, risk: C) disrupts the binding site for the TF Prdm15. (C) Both

the reference and the risk haplotype across the two CREs experience upregulation into adolescence. The reference haplotype changes from neutral to enhancer activity, while the risk haplotype changes from silencer to neutral activity. (D) The SNP in CRE #127 is an eQTL in which the risk allele is associated with reduced expression of *ARHGAP23* and *CDK12*, but increased expression of *FBXL20* and *PPP1R1B*[28]. The CREs show chromatin contacts with the likely promoters for *ARHGAP23*, *CDK12*, and *FBXL20* [45,49]. (E) Of potential target genes, *CDK12* undergoes upregulation into adulthood in human brain, consistent with our finding of P40 upregulation at this regulatory locus (boxplot from the BrainSeq Phase 2 Developmental Expression Browser[46]).

A final regulatory locus of interest in our assay contained CREs #69 and 70 on chromosome 7. Interestingly, both CREs at this locus showed generally higher expression from the risk allele than from the reference allele. Both CREs were represented by a range of barcodes and no barcode appeared to dominate the activity for either CRE (Fig. S3C). In CRE #69, the risk allele induced age-independent enhancer activity, while the reference allele showed upregulation from P10 to P40. In CRE #70, the risk allele showed upregulation at P40, while the reference allele showed constitutive repressor activity (Fig. 6A). TF motif analysis showed a number of binding motifs altered at the SNPs in both CREs. CRE #69 contains the SZ risk SNP rs9450822 (reference: G, risk: A). At rs9450822, the reference allele creates a binding site for the activating[47] TF Sox1 and the bidirectional[47] TF Fubp1, while the risk allele creates binding sites for the activators[47] Po2f1, Po2f2, and Meis2. In CRE #70, the tested risk SNP is rs4235834 (reference: G, risk: T). The reference allele creates binding sites for the repressive[47] TF Bcl11b and bidirectional[47] TFs Runx1 and Runx2. These binding sites are broken by the risk allele (Fig. 6B). The cumulative regulatory effect across the locus showed upregulation into adolescence in both haplotypes. The reference haplotype showed weak silencing at P10 which was lost at P40, while the risk haplotype showed weak enhancer activity at P10 which was strongly upregulated

during adolescence (Fig. 6C). The tested SNPs in both CREs are eQTLs in which the risk allele is associated with increased expression of nearby gene *HTR1E*[28], consistent with our finding of increased transcriptional activity from the risk haplotype at this locus (Fig. 6D). In human brain, *HTR1E* undergoes upregulation during adolescence and adulthood[46] (Fig. 6E). These CREs also show chromatin contacts in human neurons for the putative promoter of nearby gene *CNR1*[45,49] (Fig. 6D). It is possible that the risk haplotype at CREs #69 and 70 contributes to altered regulation of either *HTR1E* or *CNR1*. These hypotheses should be rigorously tested with future experimental work.

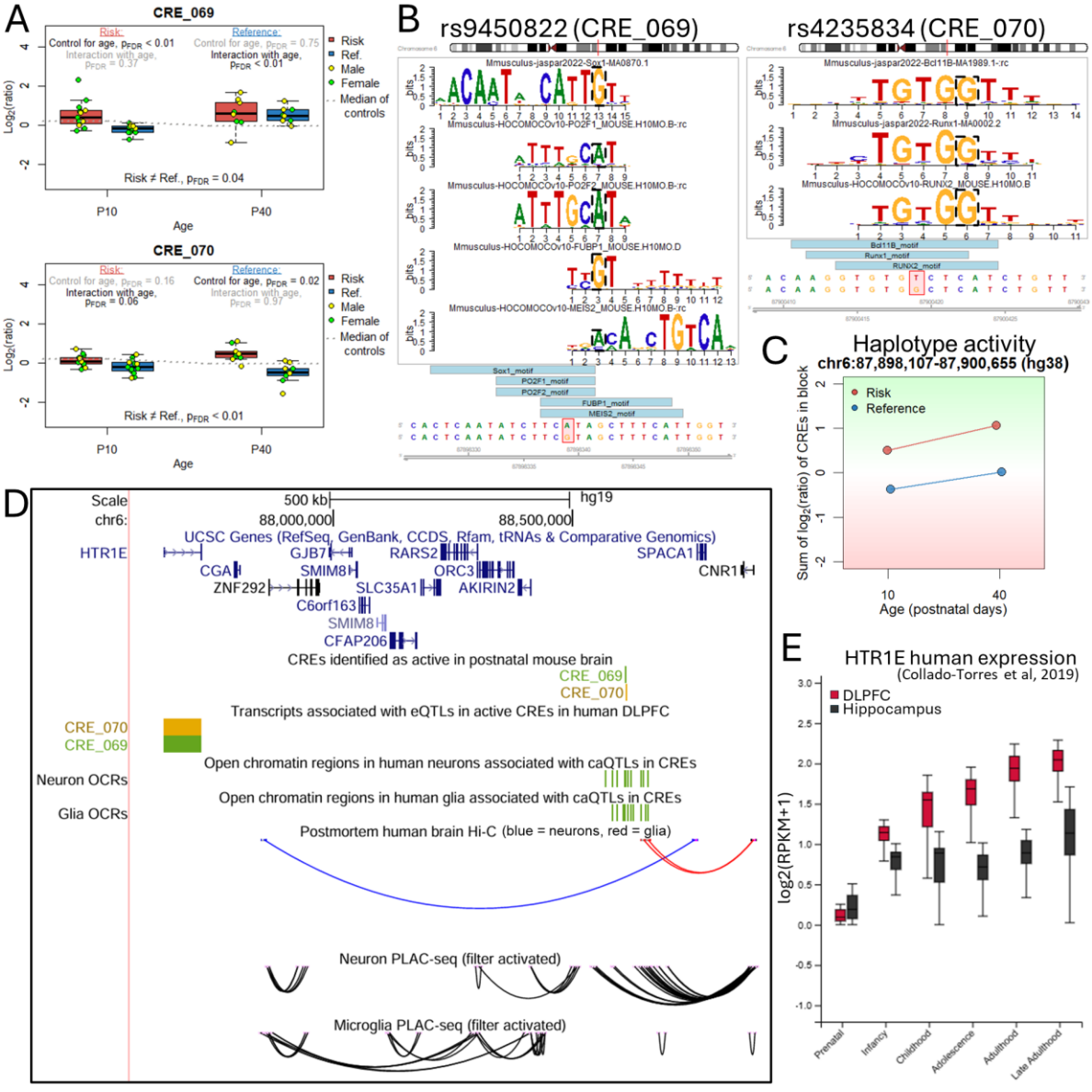


Fig. 6: Two adjacent CREs on chromosome 6 show upregulation with SZ risk alleles. (A) CREs #69 and 70 were both active in our assay and showed allelic differences. With the reference allele, CRE #69 showed upregulation of activity from P10 to P40, while the risk allele induced constitutive enhancer activity. At CRE #70, the reference allele induced constitutive repressor activity, while the risk allele showed upregulation in activity from P10 to P40. (B) TF motif analyses identified several TFs whose binding sites are broken by the SNPs in CRE #69 (SNP: rs9450822, reference: G, risk: A) and CRE #70 (SNP: rs4235834, reference: G, risk: T). (C) Across the two CREs, both the reference and risk haplotypes showed upregulation in transcriptional activity from P10 to P40. The risk haplotype acted as an enhancer at both ages, but increased in strength at P40, while the reference haplotype changed from weak silencer to relatively neutral activity. (D) The SNPs in both CREs are eQTLs in which the risk haplotype is associated with increased expression of *HTR1E*[28]. They are also caQTLs associated with local chromatin accessibility in neurons and glia[27]. These CREs show chromatin contacts with the putative promoter for *CNR1*[45,49]. (E) Of likely gene targets, *HTR1E* shows a strong pattern of upregulation into adolescence and adulthood in human brain (boxplot from the BrainSeq Phase 2 Developmental Expression Browser[46]).

Discussion

In this study, we tested candidate CREs containing SZ risk variants for regulatory activity in juvenile and adolescent mouse brain. Comparing these to a novel library of 22 empirically validated negative controls as well as empty plasmid, we identified 39 elements that were active for enhancer or silencer activity at $p_{FDR} \leq 0.10$. CREs showed reproducible activity across mouse replicates. TF motif analyses offered insights as to mechanisms which may have mediated adolescent-specific changes in activity as well as allelic differences in activity. Adolescent-specific enhancers were enriched for binding sites for the TFs Jund, Nfix, and

Znf266 (mouse ortholog Zfp266). Zfp266 is predominantly a repressor[31], while Nfix has bidirectional effects on transcription[32]. Both TFs undergo strong downregulation in adolescent mice[38], providing additional evidence for postnatal changes in activity trajectory at these CREs. Of active CREs, 19 showed FDR-significant activity differences between the SZ risk and reference alleles. These CREs were enriched for binding motifs for Emx2, Gsx2, Myod1, and Znf667. Emx2 has been strongly associated with SZ[36,37], hinting at possible convergent mechanisms between Emx2 disruption and our allelic CREs which contained binding sites for Emx2. We additionally intersected our findings with publicly available datasets on eQTLs[28], caQTLs[27], chromatin contacts[45,49], and gene expression[46] to identify potential target genes and generate hypotheses for future research. The gene targets we identified as potential candidates have been associated with SZ with a range of strength and clarity. We generate these hypotheses as recommendations for future testing to better link regulatory region SZ variants to potential mechanisms of psychiatric disease.

We first highlighted a locus on chromosome 5 with a potential regulatory link to nearby genes *GRIA1*, *FAM114A2*, and *GALNT10*. *GRIA1* encodes the AMPA receptor GluR1 subunit. *GRIA1* and GluR1 have been found to be reduced in postmortem brain from humans with SZ[50–52], and aberrant glutamatergic signaling is currently a top hypothesized risk mechanism for SZ[18]. *FAM114A2* was identified as a prioritized SZ risk gene in the most recent PGC GWAS of SZ[18], though its function is not well understood. *GALNT10* is involved in *O*-glycosylation and is underexpressed in postmortem DLPFC from humans with SZ[36]. The CREs identified at this locus may be involved in dysregulation of any of these genes in SZ. This hypothesis should be robustly tested with future research.

We also identified two CREs on chromosome #17 which showed evidence for possible regulation of nearby genes *ARHGAP23* and *CDK12*. *ARHGAP23* is an activator for Rho GTPases, which are known to be critical in neurodevelopment and have been associated with

neurodevelopmental disorders, including SZ[53,54]. *ARHGAP23* itself is significantly underexpressed in postmortem DLPFC from humans with SZ[5]. *CDK12* encodes a cyclin-dependent kinase which regulates the activity of RNA polymerase II[47]. It is critical in the maintenance of adult neuronal health[55] and has been identified as differentially expressed in peripheral blood mononuclear cells from humans with SZ[56].

Finally, we highlight a locus with two adjacent allelic CREs in which the SZ risk haplotype which may aberrantly upregulate nearby genes *HTR1E* or *CNR1*. *HTR1E* encodes the 5-HT1e serotonin receptor. It is strongly upregulated in human cortex and hippocampus during adolescence[15,57] and is overexpressed in postmortem DLPFC from humans with SZ[36]. Antipsychotic medications including clozapine and ziprasidone have described activity at the 5-HT1e receptor[58], in addition to more prominent activity at other dopamine and serotonin receptor subtypes. *HTR1E* has been shown to be hypomethylated in blood in subjects with SZ versus healthy controls[59], which may indicate a loss of silencing on *HTR1E* in these subjects. *CNR1* encodes the CB1 cannabinoid receptor. *CNR1* has repeatedly been associated with SZ, but comparative expression analyses between SZ and control brain have not identified a consistent direction of change. One study found that *CNR1* was overall reduced in SZ DLPFC, but increased in PFC of SZ subjects who died by suicide[60]. Another study found that the *CNR1* promoter was hypomethylated in peripheral blood from humans with SZ, indicating a likely loss of silencing[61]. The same study identified upregulation of *CNR1* in prefrontal cortex in a rat model of SZ[61]. Interestingly, use of antipsychotic medication is associated with downregulation of the CB1 receptor[62]. While this may hint at the relevance of dysregulated *CNR1* to SZ, it also substantially complicates the ability to study *CNR1* in postmortem tissue from humans with SZ. As such, more work will be necessary to determine the relationship between *CNR1* and SZ.

As well as identifying CREs active in postnatal mouse brain, we developed two novel resources which may be useful to the MPRA community. First, we designed and cloned a library of

barcoded MPRA plasmids which are optimized to minimize potential confounding effects in MPRAs. Second, we developed and validated a library of 22 negative controls which did not show regulatory activity versus empty plasmid in postnatal mouse brain. Negative controls are critical for rigorous MPRAs, but are not always used due to a lack of validated candidates. We share these resources with other researchers and report all barcodes and negative controls in Supplementary Tables 2 and 3.

Inclusion of negative controls in our experiments allowed us to model the background transcription rate from our MPRA construct and subsequently identify a number of silencer elements in addition to enhancers. Historically, MPRAs have focused their attention on testing enhancer elements, which are typically identified as the elements showing the highest rates of transcription[4,24]. However, silencers are pervasive in the noncoding human genome[63] and have been robustly identified via MPRA in those experiments which have tested for them[64,65]. Interestingly, many elements have been found to switch between enhancer and silencer activity[66], and silencers have been widely identified at supposedly canonical epigenetic enhancer marks including H3K4me1 and H3K27ac[64,66]. Silencer disruption has been gaining recognition as a mechanism of disease risk in SZ as well as in other psychiatric and neurological disorders[6]. These data highlight the need for increased characterization of silencer activity. We identified in our assay a number of silencer elements, including elements whose silencing function is induced or disrupted in the presence of the SZ risk allele.

We are also, to our knowledge, the first to deploy a longitudinal MPRA spanning preadolescent and adolescent development. The temporally dynamic activity of CREs is well documented and particularly well characterized in the context of prenatal development[67]. Various methods for postnatal *in vivo* MPRA exist[24,68–70], but remain challenging. As such, MPRAs are more often conducted *in vitro*[4,71–73]. However, *in vitro* systems cannot model the full contexts of developmental processes, experience, and interacting biological systems which *in vivo* models

are better equipped to capture. Particularly in the case of adolescent and adult-onset disorders like SZ, capturing the neurodevelopmental processes that occur during adolescence is critical for our understanding of how disrupted biological processes contribute to disease. We identified several CREs which showed altered activity between juvenile and adolescent brain development in our assay. These experiments may lay the groundwork for postnatal MPRAs in adolescence and later stages of development and aging, offering the potential to dissect regulatory activity of adult-onset diseases.

Our work features limitations. Large-scale fine-mapping efforts have made strides in identifying the most likely causal signals underlying GWAS associations to SZ[18,74]. However, given the low thresholds for linkage disequilibrium typically used to prioritize loci for these studies (often R^2 = 0.1)[18,74], it is possible that these efforts miss some independently acting causal risk loci. As such, we did not choose to use fine-mapped SZ datasets as a filtering criterion when selecting variants to test. However, this means that we are likely missing some of the best-evidenced risk loci in our assay, and possibly testing CREs which are associated with SZ only via their proximity to the true causal signal. While this enabled us to test potentially independently causal loci which are in high LD with finemap-prioritized risk variants, our CREs require further research to evaluate their likelihood of causal association to SZ. We also did not capture sufficient coverage of both alleles to merit testing of allelic differences for all CREs in our assay. Consequently, CREs which were not tested for allelic differences may actually show strong allelic differences which we have not reported. We focus our analyses on those CREs which did show allelic differences, and report all other significant results in Table 1, including whether allelic effects were tested for each CRE. As another limitation to our assay, the relevance of our findings rests on the assumption that TF activity is comparable across postnatal human and mouse brain, and particularly within the context of adolescent neurodevelopment. Both brain TF activity and neurodevelopmental milestones are regarded as generally well conserved across

human and mouse[75–79]. This has formed the foundation for *in vivo* MPRAs as well as the use of mice in general to model phenomena of relevance to human adolescence. However, there are likely to be differences of relevance to our CREs that we are not equipped to decipher with the described experiments. For all TFs which we highlighted as disrupted by SZ risk SNPs in our assay, we verified conservation of expression in both postnatal mouse and adult human cortex. Still, our identified active CREs should be replicated in human neuronal lines. As an ectopic assay, we also do not capture the native chromatin context around our CREs and instead our results reflect only the binding of available TFs. This misses much of the critical context that modulates CRE activity in the nucleus. Again, our experiments should be repeated in human cell lines using methods such as perturb-seq[80] to better capture the effects of the local chromatin conformation. We also do not examine cell-type specificity of regulatory activity in our assay. It is likely that CREs affected by SZ risk SNPs show differential activity across neuronal and glial subtypes; combining regulatory assays such as MPRAs with single-cell RNA sequencing or fluorescence-activated cell sorting may help deconvolute these differences across cell types. Future work should intersect the data from our experiments with data generated by these suggested methods in order to best characterize CREs which may show aberrant activity in human adolescents at risk of developing SZ. Finally, we generate hypotheses for genes targeted by our CREs based on intersection of our findings with publicly available datasets. We note that our experiments only tested for activity of CREs in an ectopic assay and explicitly do not test for associations between CREs and gene targets. Any gene targets we suggest will require robust experimentation. Further, these genes are regulated by a much larger network of *cis*- and *trans*-regulatory elements than were tested in our assay, and as such we do not capture the entire regulatory context acting on these genes. However, we generate these hypotheses in an effort to help prioritize a feasible set of loci out of the thousands of risk variants which have been identified in SZ GWASs. We hope these hypotheses will support future research into causes, preventions, and treatment targets in SZ.

Despite these limitations, our study successfully identified a set of 39 active CREs in postnatal mouse brain which contain prioritized SZ risk-associated variants. 19 of these showed differential activity associated with the SZ risk versus reference allele. CREs which underwent developmental changes in activity were enriched for TFs with postnatally dynamic changes in expression. CREs which showed allelic differences were enriched for a number of TF motifs, including for the SZ-associated Emx2. We highlight three regulatory regions from our MPRA for detailed analysis. The allelic differences we identified may underlie our tested variants' GWAS associations with SZ. Developmental changes in CRE activity may contribute to the adolescent-onset nature of SZ symptoms. We generated hypotheses for gene targets which may link these noncoding variants to mechanisms of SZ. Our experiments also contributed to the expansion of MPRA utility by modeling an adolescent longitudinal *in vivo* MPRA. We additionally developed and share a library of optimized MPRA barcoded plasmids as well as an empirically validated library of MPRA negative controls for use in mouse brain. We hope these resources will be useful for future MPRAs and enable rigorous testing of CREs *in vivo*. The CREs we identified using these methods should be replicated in human models, and our generated hypotheses for potential gene targets should be rigorously tested. However, these efforts may help map noncoding SZ risk variants onto mechanisms of disease, and hopefully contribute to improved identification of postnatal targets for prevention and treatment in SZ.

Methods

Regulatory committee approvals

Human DNA collection and use were approved by the University of California, Davis, Institutional Review Board and all subjects gave written consent and were paid for their participation. Mouse experimental work was approved by the UC Davis Institutional Animal Care and Use Committee (Protocol #22339).

MPRA barcode library design

We designed a library of MPRA barcodes containing up to 409,600 unique sequences of 25 bp in length. To do this, we generated random "scaffolds" of 25 bases, containing seven wildcards per scaffold (e.g. NCGNATCNGCTNCANCGTNACNATC where Ns represent wildcard nucleotides which will be replaced by A, C, T, and G in equal portion during barcode synthesis). In R, we generated all possible sequences that would result when barcodes were cloned into our MPRA plasmid and wildcard nucleotides were replaced with As, Cs, Ts, and Gs. To minimize the chance for barcode-induced confounds in our assay, we then filtered out any scaffold sequences that would produce barcodes containing binding motifs for 244 known eukaryotic RNA binding proteins[26], single-nucleotide repeats longer than three bases, or overly high or low GC content. We repeated this process until we had identified 25 scaffold sequences which passed these parameters. We ordered a pool of synthesized oligonucleotides composed of our scaffold sequences from Integrative DNA Technologies (Coralville, IA). During synthesis, all Ns in our scaffold sequences were replaced with As, Cs, Ts, and Gs, resulting in a pool of up to 409,600 unique sequences (4 nucleotides ^ 7 wildcards = 409,600). We cloned this pool into the 3' UTR of EGFP in our MPRA plasmid as described in "Batch cloning protocol" below.

Negative control candidate selection

We developed a library of negative controls containing both empirically determined negative control candidates as well as *in silico* predicted negative controls approximately 900 bp in length. For the empirically determined library, we selected 15 elements in the third intron of the human gene *CACNA1C* which did not show enhancer activity in a previous MPRA conducted in postnatal mouse brain[24]. These elements were PCR amplified out of genotyped DNA obtained from healthy human subjects[29]. For the *in silico* predicted library, ten elements were selected at random from the mouse genome based on a lack of evidence for regulatory activity in mm10 using the UCSC Genome Browser[81]. These were specifically screened for a lack of ENCODE

consensus CREs[82], conservation across vertebrates, ATAC peaks in embryonic forebrain[82], NCBI RefSeq consensus functional elements[83], H3K4me3 in embryonic brain[82], H3K27ac in embryonic brain[82], and repeat elements[84]. These elements were PCR amplified out of DNA obtained from wild-type C57/BL6J mice. Our candidate negative control library consisted of a final total of 25 elements.

CRE candidate selection

Autosomal summary statistics were downloaded for the most recent Psychiatric Genomics Consortium (PGC) GWAS meta-analysis for SZ[18]. P-values were FDR-adjusted using the Benjamini & Hochberg (1995)[85] procedure. Odds ratios were calculated from reported β-values. Variants were filtered to those with $p_{FDR} \leq 0.10$, odds ratios ≥ 1.02 or ≤ 0.98, and where both the variant association and its direction of effect were consistent in the replication studies reported the GWAS summary statistics. We next developed a gene set of synaptically relevant genes by merging the GoSlim synapse gene set from QuickGO[86] with four previously published gene sets featuring genes relevant to glutamatergic, GABAergic, dopaminergic, and serotonergic synapses[29]. Annotations containing the following GoSlim terms were removed: “skeletal”, “neuromuscular”, “synapse”, “acetylcholine”, “calyx of Held”, “cerebellar”, “cochlear”, “cholinergic”, “cone cell pedicle”, “photoreceptor”, “ribbon synapse”, and “innervation”. We subsetted human DLPFC eQTL data from Fromer et al. (2016)[28] to eQTLs associated with altered expression of genes within our newly developed gene set. We further filtered these eQTLs to those that had also been identified as caQTLs in human DLPFC[27] (accessed March 2023 via preprint). We intersected this list of dual eQTL-caQTL variants with our list of prioritized GWAS variants, identifying 299 which overlapped. We then queried variant impacts using the Ensembl Variant Effect Predictor tool[87], and subsetted our variant list to those that were identified as "downstream_gene_variant", "intron_variant", "upstream_gene_variant", or "intergenic_variant". This left 152 variants which were divisible across 139 candidate CREs of

300-500 bp in length. We discarded six CREs that contained repeat regions which complicated primer design, and three additional CREs that were determined to be too close to a gene exon. Our final list consisted of 130 candidate CREs containing prioritized SZ GWAS variants. Using genotyped DNA obtained from healthy human subjects[29], we identified individuals who were homozygous for either the reference or risk haplotypes across large LD blocks spanning CREs in our assay. We PCR amplified CREs from identified DNA samples and proceeded with cloning procedures as described below.

Batch cloning protocol

Each round of cloning (separately for barcodes, negative controls, and CRE library) was conducted on several large LB agar plates using *E. coli* under antibiotic selection with carbenicillin, for which our MPRA cloning plasmid contains a resistance gene. We used the InFusion cloning kit (Takara # 638948) and NEB Stable Competent *E. coli* (New England Biolabs, # C3040H), following the manufacturers' protocols. To ensure library diversity was achieved during cloning, we used Azenta Life Sciences (South Plainfield, NJ) to perform Sanger sequencing on a subset of colonies, and otherwise combined the rest of the colonies from all plates into a flask of LB. The pool of colonies was maxiprepped using the Qiagen HiSpeed Plasmid Maxi Kit (Qiagen # 12662) following the manufacturer's protocol. Finally, we performed a restriction digest using XmaI (New England Biolabs # R0180S) on a sample from each maxiprep to ensure integrity of the ITRs. This cloning protocol was conducted first to clone the MPRA barcode library into our plasmid backbone. The plasmid from this maxiprep was then used for separate cloning of negative control and experimental CRE libraries. For our negative control library, we additionally cloned in a small portion of empty plasmid into which we did not insert a candidate CRE, which enabled us conduct the negative control validation experiments versus empty plasmid described in "Results".

Matching CREs and barcodes

To associate regulatory elements to barcodes, long-read sequencing of our final cloned plasmid libraries was performed by Plasmidsaurus (Eugene, OR) using Oxford Nanopore Technology. Using the stringdist package[88] in R, we identified sequencing reads which contained our CRE and negative sequences. We allowed for up to a 5% error rate in base matching based on Levenshtein distance[89], given current error rate estimations for Nanopore sequencing[90]. For our empty plasmid sequences included for the negative control experiments, we used the sequence spanning the cloning site from our MPRA plasmid backbone. For our CRE library, we separated reads into those which better matched the reference or risk allele at the risk SNP. We next identified barcode sequences by extracting the 25 bases which were flanked by the plasmid backbone sequence in the 3' UTR barcode cloning site. To identify unique barcodes, we adapted methodologies from McAfee et al (2023)[4]. We first identified the most frequently read barcode per CRE, considering risk and reference alleles as two distinct CREs. We then used stringdist[88] to identify all barcodes which were matched to the same CRE and appeared to be misread duplicates of this most frequent barcode, allowing up to two misread bases. We added the read counts for these presumed duplicates to that of the most frequent barcode, and subsequently removed these sequences from the read list. We then repeated this process for the second most frequent barcode, and then moved onto the third most frequent barcode. This process was continued until only the least frequent barcode which had not been matched to others was remaining. Upon developing these barcode lists for each CRE, we next filtered to ensure that barcodes were not duplicated across multiple CREs. We first removed any barcode that was represented in both the negative and the experimental CRE libraries. Next, within the experimental CRE library and separately within the negative control library, we identified remaining barcodes which had been matched to multiple CREs. If the read count for a barcode was 50x higher in its most frequently read CRE than the sum of its reads associating it to all other CREs in the library, we retained that barcode as representing only the highest-matched

CRE. For example, if a barcode showed 100 reads for CRE #1, one read for CRE #2, and one read for CRE #3, that barcode was identified as 50x more prevalent in CRE #1 than the sum of its other associations, and as such was retained as a barcode representing CRE #1. In theory, this should enable detection of the correct CRE from this barcode in our assay 98% of the time. Only barcodes with at least two reads supporting their CRE association were retained for MPRA data analysis.

Viruses and injections

Both the negative control library and the experimental CRE library were packaged into separate AAVs of serotype PHP.eB[91]. The CAG-mRuby3 plasmid used for IHC experiments was also packaged into AAV-PHP.eB. All viruses were packaged at the Stanford Gene Vector and Virus Core (RRID:SCR_023250). A total of 2 µl of injectate was delivered via intracerebroventricular (ICV) injections to P0 mice at 1 µl per hemisphere, following procedures outlined in Warren et al (2022)[68]. For negative control experiments, the negative control AAV was injected into 11 mice at 2.10×10^{10} viral genomes (vG) per animal (final titer 1.05×10^{13} vG/ml). For CRE experiments, the CRE experimental AAV and negative control AAV were pooled in a 3:1 vG ratio, and 3.2×10^{10} total vG (CRE AAV: 2.40×10^{10} vG, final titer 1.20×10^{13} vG/ml. Negative control AAV: 8.00×10^{9} vG, final titer 4.00×10^{12} vG/ml) were injected via ICV injections to P0 mice at 1 µl per hemisphere. Twelve mice were injected for P10 sequencing experiments, and eight mice were injected for P40 sequencing experiments. For imaging-endpoint experiments, the CRE experimental AAV and CAG-mRuby3 AAV were pooled in a 443:1 vG ratio (CRE AAV: 2.22×10^{10} vG, final titer 1.11×10^{13} vG/ml. CAG-mRuby3 AAV: 5.00×10^{7} vG, final titer 2.50×10^{10} vG/ml). Mice were injected at P0 and collected at P10 and P40 for processing as described in “Immunohistochemistry and fluorescence microscopy”.

Immunohistochemistry and fluorescence microscopy

Mice were anesthetized with isoflurane and transcardially perfused at P10 and P40 for imaging-endpoint experiments. Brains were dissected out and submerged in 4% PFA in PBS overnight and transferred to 30% sucrose in PBS the next day for an additional 3-4 days. Brains were then frozen in Tissue-Tek O.C.T. Compound (Sakura # 4583), cut into 30 µm sections on a cryostat, and collected into PBS. Sections were stained for GFP and NeuN following the IHC protocol described in Warren et al (2022)[68]. Primary antibodies included chicken anti-GFP (ThermoFisher #A10262, 1:1000) and rabbit anti-NeuN (Abcam #ab104225, 1:1000). Secondary antibodies included a Donkey anti-chicken AlexaFluor-488 (Jackson ImmunoResearch #703-545-155, 1:500) and a Donkey anti-rabbit AlexaFluor-647 (Jackson ImmunoResearch #712-605-152, 1:500). Confocal images were acquired on a Zeiss LSM800 microscope using a 0.25 NA 5x air objective. Images for P10 and P40 brains were acquired during the same imaging session and using the same settings for laser intensity, PMT gain, and pixel dwell time. Background was subtracted and contrast was adjusted using Fiji[92].

RNA/DNA extractions and sequencing

Mice were anesthetized with isoflurane and sacrificed via cervical dislocation followed by decapitation. Dorsal cortex and hippocampus were dissected on ice and immediately submerged into RNAlater (ThermoFisher # AM7020). Samples were stored at 4 °C for a maximum of one week. DNA and RNA were extracted from the same tissue samples using the Qiagen AllPrep DNA/RNA Purification Mini kit (Qiagen # 80204). One modification was made to the protocol: to ensure we obtained a homogenous sample representing the entirety of collected tissue, the full sample tissue sample was homogenized in lysis buffer, rather than a maximum of 30 mg as recommended by the protocol. The volumes of RLT plus and β-mercaptoethanol were scaled up accordingly. A sample of only 1.5 ml of the homogenized solution was used for the

following centrifugation step. Subsequently, 700 µl of supernatant was added to the DNA spin column, and the rest of the procedure was completed following the manufacturer's protocol.

A maximum of 2.5 µg of RNA per sample was treated with ezDNase (Invitrogen # 11766051) following the manufacturer's protocol. To enrich for transcripts corresponding to our MPRA barcodes during reverse transcription, a maximum of 1.25 µg of RNA was mixed with 1 µl of reverse primer at 10 µM which would bind to our MPRA plasmid and nuclease-free water up to a total volume of 8 µl. This solution was incubated at 65 °C for five minutes followed by 4 °C for one minute. We then added 2 µl of Superscript IV VILO Master Mix (Invitrogen # 11756050) and proceeded with reverse transcription per the manufacturer's protocol. We then incubated the subsequent cDNA with RNAse H (New England Biolabs # M0297S) at 37 °C for 25 minutes to remove RNA bound to cDNA molecules.

For both the originally extracted DNA and the newly created cDNA, we PCR amplified the barcode region for sequencing. Reverse primers contained unique, predefined barcode sequences six bases in length. Given that AAV MPRA protocols often result in sparse transduction and only a small portion of DNA obtained from the whole tissue actually contains the template to be amplified, the amount of template used for PCR needs to be scaled up substantially for detection at rates comparable to those obtained from PCR using native cellular DNA. As such, 250 ng of DNA and cDNA were loaded into each 25 µl PCR. We used CloneAmp PCR MasterMix for all PCRs, following the manufacturer's protocols. PCRs ran for 35 cycles. PCR products were then run on 1% agarose gels at 85 V for 30 minutes. Bands corresponding to the expected PCR products were verified for each sample and extracted from the gels using the NucleoSpin Gel and PCR Clean-Up kit (Takara # 740609.50). Samples were cleaned an additional time on the Zymo DNA Clean & Concentrator-25 kit (Zymo Research # D4033). Cleaned, PCR-barcoded samples were pooled and submitted for Illumina amplicon sequencing performed by Azenta Life Sciences (South Plainfield, NJ).

CRE identification in sequencing results

Sequences surrounding the MPRA barcode and PCR barcode were identified in each sequencing read. Given high fidelity with Illumina sequencing[93], no misread bases were permitted. Segments corresponding to the MPRA and PCR barcodes were extracted. Reads were assigned to their original samples based on exact matches to PCR barcodes. MPRA barcodes were then matched to our previously developed key of MPRA barcodes based on exact matches to MPRA barcodes. To ensure resolution to calculate RNA/DNA ratios, all barcodes with fewer than five reads in a given animal's DNA sequencing results were removed from that animal's DNA and RNA sequencing results. Barcodes which were identified in an animal's DNA sequencing at ≥ 5 reads, but which were not identified in that animal's RNA sequencing, were padded with zeros in the RNA. Barcodes were summed per CRE. CRE counts were then normalized as a proportion of total sequencing reads for that animal's RNA and DNA, such that if a CRE was identified with 100 reads in an animal's DNA, and a total of 10,000 reads had been returned for that animal's DNA, that CRE's DNA count proportion was 100 ÷ 10,000 = 0.01. To avoid divisions and logs with zeros, +1 was added to both the numerator and denominator during this proportion calculation step. $\text{Log}_2(\frac{RNA}{DNA})$ was calculated per CRE, per mouse. To minimize potential confounds or false positives/negatives due to low technical replicates (barcodes), we only retained CREs represented by at least 5 barcodes in a given mouse. Finally, we removed CREs which were represented by fewer than 50 reads in a mouse's DNA.

Negative control experiments

Mice injected with the negative control virus were collected on P28. DNA and RNA collection, DNA and RNA sequencing, and CRE identification were conducted as described above. To control for potential confounding effects of covariates, we regressed $\log_2(\frac{RNA}{DNA})$ against both sex

and number of barcodes for all CREs and proceeded with the residuals from this regression model for all analyses. For simplicity, we refer to this regressed value as "$\log_2(\frac{RNA}{DNA})$" for the remainder of the text. Separately for each negative control, we modeled $\log_2(\frac{RNA}{DNA})$ against experimental condition (negative control vs. empty plasmid, coded as 1 or 0) ($\log_2(\frac{RNA}{DNA})$ ~ condition). 24 negative candidates were detected sufficiently for testing, and two of these were significantly different in activity from empty plasmid at an unadjusted p-value threshold of 0.05. The remaining 22 candidates and the empty plasmid sequence were retained as validated negative controls for subsequent analyses.

Allelic analyses

We first tested CREs for allelic differences. As in the negative control experiments, we regressed $\log_2(\frac{RNA}{DNA})$ against both sex and number of barcodes for all CREs, including controls, and proceeded with the residuals from this regression model for all analyses. For simplicity, we refer to this regressed value as "$\log_2(\frac{RNA}{DNA})$" for the remainder of the text. To capture alleles which may show consistent differences in activity across ages, we modeled $\log_2(\frac{RNA}{DNA})$ against allele, coded as a binary 0 or 1, and controlling for age ($\log_2(\frac{RNA}{DNA})$ ~ allele + age). To capture alleles which may show different trajectories of activity across ages, we modeled $\log_2(\frac{RNA}{DNA})$ against allele and included age as an interaction effect ($\log_2(\frac{RNA}{DNA})$ ~ allele*age). We FDR-adjusted all p-values together using the Benjamini & Hochberg (1995)[85] procedure. For all subsequent analyses, we treated alleles as distinct CREs if they showed significant differences either controlling for age or as an interaction effect with age at $p_{FDR} \leq 0.10$. For all other CREs, alleles were merged and count proportions across alleles were summed. $\text{Log}_2(\frac{RNA}{DNA})$ calculations and

regression analyses were repeated for the entire library after merging CREs without allelic differences. These regressed $\log_2(\frac{RNA}{DNA})$ values were used for all final analyses.

Correlations and concordance

All calculations for concordance and correlation were conducted separately within P10 mice and within P40 mice. Spearman's correlations for $\log_2$(RNA), $\log_2$(DNA), and $\log_2(\frac{RNA}{DNA})$ are reported across mice. For concordance, we assigned a CRE as "upregulated" or "downregulated" in a given mouse if that CRE's $\log_2(\frac{RNA}{DNA})$ was respectively +1 or -1 away (representing a doubling or halving of activity) from the median of negative controls in all mice at that age. We removed a CRE for a mouse if it was not assigned as either upregulated or downregulated via these criteria. For all CREs retained in at least two mice, we calculated how many mice received the same assignment (upregulated vs. downregulated) out of the total number of mice in which that CRE was retained. For example, if a CRE was retained in 5 mice, and was assigned as downregulated in 4 mice but upregulated in 1 mouse, then 4 ÷ 5 = 80% concordance. To calculate the null distribution, we took the range of $\log_2(\frac{RNA}{DNA})$ which had been identified in our CREs and generated 1,000 random numbers within this range. We repeated this for all CREs, so that each CRE at each age was assigned 1,000 observations of $\log_2(\frac{RNA}{DNA})$ which fell within the empirically observed range. We then repeated the described concordance calculation procedure. To determine whether the observed concordance was greater than the null set, we performed a one-tailed Wilcoxon sign rank test. We calculated and report kernel densities for the observed and null distributions of concordance using the density() function in R with the bandwidth value set to 0.1.

Identifying active CREs

To test for activity, we required that CREs be detected in at least three mice at each P10 and P40. For each CRE which was detected sufficiently at both ages, that CRE's observations in all mice were appended to a data table along with all negative controls, including empty plasmid, in all mice. We then modeled $\log_2(\frac{RNA}{DNA})$ against experimental condition (CRE or control, coded as 1 or 0), controlling for age ($\log_2(\frac{RNA}{DNA})$ ~ condition + age). To capture CREs whose trajectories changed over time, we also repeated this modeling while including age as an interaction effect with condition ($\log_2(\frac{RNA}{DNA})$ ~ condition*age). Alleles which had been identified as significantly different at $p_{FDR} \leq 0.10$ were calculated as separate CREs. P-values were pooled across all tests and FDR-adjusted using the Benjamini & Hochberg (1995)[85] procedure. CREs were identified as active if they were significantly different from controls either controlling for age or as an interaction effect with age at $p_{FDR} \leq 0.10$.

Haplotype activity

For haplotype activity plots reported in the text, we first calculated the median activity reported per CRE, per age, across all mice in which the CRE was identified at that age. We summed these median activities across active CREs within the wider regulatory locus, within-haplotype, within-age. For example, if the risk allele at CRE #1 had a $\log_2(\frac{RNA}{DNA})$ of 1.2 at P10, and the risk allele at CRE #2 had a $\log_2(\frac{RNA}{DNA})$ of -0.5 at P10, then the summed value at P10 for the risk haplotype spanning that regulatory locus was 1.2 + (-0.5) = 0.7.

TF motif enrichment analyses

To test for TFs which may mediate longitudinal changes at adolescent-specific enhancers, we used the Simple Enrichment Analysis (SEA)[94] tool from MEME Suite[95]. We looked for TF motif

enrichments from the HOCOMOCOv12 TF dataset[96], using sequences which were neutral in our assay (unadjusted p > 0.05) as controls. We retained default settings for all test parameters. We repeated this analysis for active CREs which showed allelic differences, again using neutral sequences as controls. For enriched TFs, we report binding motif logos from JASPAR[97] and expression trajectories of *Nfix* and *Zfp266* from the Evo-Devo Sex-Biased Mammalian Genes app[38] in the text.

For allelic CREs in Figures 4-6, we used the motifbreakR package[98] in R to identify mouse-expressed TFs from HOCOMOCOv10[96] and JASPAR 2022[99] whose binding sites were predicted to be disrupted by the risk SNP (using human sequence from hg38). We retained only those TFs which were predicted to bind at a nominal p-value threshold of 1e-04 and whose effect size for disrupted binding by the SNP was categorized as “strong”. We then verified expression of predicted TFs in P14 and/or P28 mouse cortex using the Allen Developing Mouse Brain Atlas[100] and only retained those which were identified as expressed. Data was unavailable on the Allen Developing Mouse Brain Atlas for a small portion of TFs. For these, we verified expression at Theiler stage[101] 28 in mouse cerebral cortex via the Mouse Gene Expression Database database[102]. We finally filtered TFs to those that were also expressed in adult human cortex using GTEx[35]. Using this final list, we generated the TF plots reported in the paper using mouse TFs from HOCOMOVOv10[96] and JASPAR 2022[99].

Acknowledgements

Human subject recruitment and blood draws were supported by NIMH grant R01MH059883. Human DNA extraction and genotyping were supported by the UC Davis-HKU Collaboration Grant. MPRA cloning, experiments, analyses, and writing were supported by the UC Davis LaMP T32 (NIMH T32MH112507), the NIMH NRSA F31MH129135, and NIGMS grant R35GM119831.

References

1. Roussos, P. *et al.* A Role for Noncoding Variation in Schizophrenia. *Cell Reports* (2014) doi:10.1016/j.celrep.2014.10.015.
2. Girdhar, K. *et al.* Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nat Neurosci* **21**, 1126–1136 (2018).
3. Huo, Y., Li, S., Liu, J., Li, X. & Luo, X.-J. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat Commun* **10**, 670 (2019).
4. McAfee, J. C. *et al.* Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. *Cell Genomics* **3**, 100404 (2023).
5. Dong, P., Voloudakis, G., Fullard, J. F., Hoffman, G. E. & Roussos, P. Convergence of the dysregulated regulome in schizophrenia with polygenic risk and evolutionarily constrained enhancers. *Mol Psychiatry* **29**, 782–792 (2024).
6. Huang, D. & Ovcharenko, I. The contribution of silencer variants to human diseases. Preprint at https://doi.org/10.1101/2024.05.17.24307558 (2024).
7. Schrode, N. *et al.* Synergistic effects of common schizophrenia risk variants. *Nat Genet* **51**, 1475–1485 (2019).
8. Townsley, K. G. *et al. Convergent Impact of Schizophrenia Risk Genes*. http://biorxiv.org/lookup/doi/10.1101/2022.03.29.486286 (2022) doi:10.1101/2022.03.29.486286.
9. Addington, J. *et al.* North American Prodrome Longitudinal Study (NAPLS 2): The Prodromal Symptoms. *Journal of Nervous & Mental Disease* **203**, 328–335 (2015).
10. Abel, K. M., Drake, R. & Goldstein, J. M. Sex differences in schizophrenia. *International Review of Psychiatry* **22**, 417–428 (2010).
11. Pergament, E. The Future of Prenatal Diagnosis and Screening. *JCM* **3**, 1291–1301 (2014).
12. Hendriks, S. *et al.* Precarious hope: Ethical considerations for offering experimental fetal therapies outside of research after initial studies in humans. *Prenatal Diagnosis* **44**, 180–186 (2024).
13. Hendriks, S. *et al.* A New Ethical Framework for Assessing the Unique Challenges of Fetal Therapy Trials. *The American Journal of Bioethics* **22**, 45–61 (2022).
14. Bianchi, D. W. From prenatal genomic diagnosis to fetal personalized medicine: progress and challenges. *Nat Med* **18**, 1041–1051 (2012).
15. The BrainSeq Consortium *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**, 1117–1125 (2018).
16. Clifton, N. E. *et al.* Dynamic expression of genes associated with schizophrenia and bipolar disorder across development. *Transl Psychiatry* **9**, 74 (2019).
17. Selemon, L. D. A role for synaptic plasticity in the adolescent development of executive function. *Transl Psychiatry* **3**, e238–e238 (2013).
18. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
19. Aryal, S. *et al. Deep Proteomics Identifies Shared Molecular Pathway Alterations in Synapses of Schizophrenia and Bipolar Disorder Patients and Mouse Model*.

http://biorxiv.org/lookup/doi/10.1101/2022.09.21.508852 (2022) doi:10.1101/2022.09.21.508852.
20. Sellgren, C. M. *et al.* Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nature Neuroscience* (2019) doi:10.1038/s41593-018-0334-7.
21. Van Berlekom, A. B. *et al.* Synapse Pathology in Schizophrenia: A Meta-analysis of Postsynaptic Elements in Postmortem Brain Studies. *Schizophrenia Bulletin* (2020) doi:10.1093/schbul/sbz060.
22. Feinberg, I. Schizophrenia: Caused by a fault in programmed synaptic elimination during adolescence? *Journal of Psychiatric Research* **17**, 319–334 (1982).
23. Nord, A. S. & West, A. E. Neurobiological functions of transcriptional enhancers. *Nature Neuroscience* (2020) doi:10.1038/s41593-019-0538-5.
24. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
25. deBoer, E., Antoniou, M., Mignotte, V., Wall, L. & Grosveld, F. The human beta-globin promoter; nuclear protein factors and erythroid specific induction of transcription. *The EMBO Journal* **7**, 4203–4212 (1988).
26. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
27. Zeng, B. *et al.* Genetic regulation of cell type–specific chromatin accessibility shapes brain disease etiology. *Science* **384**, eadh4265 (2024).
28. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442–1453 (2016).
29. Warren, T. L. *et al.* Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis. *Molecular Psychiatry* (2024) doi:10.1038/s41380-024-02457-0.
30. Retallick-Townsley, K. G. *et al.* Dynamic stress- and inflammatory-based regulation of psychiatric risk loci in human neurons. Preprint at https://doi.org/10.1101/2024.07.09.602755 (2024).
31. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
32. Messina, G. *et al.* Nfix Regulates Fetal-Specific Transcription in Developing Skeletal Muscle. *Cell* **140**, 554–566 (2010).
33. Cecchi, C. Emx2 : a gene responsible for cortical development, regionalization and area specification. *Gene* **291**, 1–9 (2002).
34. Gulisano, M., Broccoli, V., Pardini, C. & Boncinelli, E. *Emx1* and *Emx2* Show Different Patterns of Expression During Proliferation and Differentiation of the Developing Cerebral Cortex in the Mouse. *Eur J of Neuroscience* **8**, 1037–1050 (1996).
35. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
36. Xia, Y. *et al.* Transcriptomic sex differences in postmortem brain samples from patients with psychiatric disorders. *Sci. Transl. Med.* **16**, (2024).
37. Li, Z. *et al.* Identification of potential biomarkers and their correlation with immune infiltration cells in schizophrenia using combinative bioinformatics strategy. *Psychiatry*

Research **314**, 114658 (2022).
38. Rodríguez-Montes, L. *et al.* Sex-biased gene expression across mammalian organ development and evolution. *Science* **382**, eadf1046 (2023).
39. Dong, J., Scott, T. G., Mukherjee, R. & Guertin, M. J. ZNF143 binds DNA and stimulates transcription initiation to activate and repress direct target genes. Preprint at https://doi.org/10.1101/2024.05.13.594008 (2024).
40. Kreimer, A. *et al.* Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat Commun* **13**, 1504 (2022).
41. Zheng, Y. *et al.* Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells. *Nature* **445**, 936–940 (2007).
42. Matsuda, T. *et al.* Pioneer Factor NeuroD1 Rearranges Transcriptional and Epigenetic Profiles to Execute Microglia-Neuron Conversion. *Neuron* **101**, 472-485.e7 (2019).
43. Chang, J. C. *et al.* Prdm13 Mediates the Balance of Inhibitory and Excitatory Neurons in Somatosensory Circuits. *Developmental Cell* **25**, 182–195 (2013).
44. Verheul, T. C. J., Van Hijfte, L., Perenthaler, E. & Barakat, T. S. The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1. *Front. Cell Dev. Biol.* **8**, 592164 (2020).
45. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* (2019) doi:10.1126/science.aay0793.
46. Collado-Torres, L. *et al.* Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron* **103**, 203-216.e8 (2019).
47. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1.30.1-1.30.33 (2016).
48. Leszczyński, P. *et al.* Emerging Roles of PRDM Factors in Stem Cells and Neuronal System: Cofactor Dependent Regulation of PRDM3/16 and FOG1/2 (Novel PRDM Factors). *Cells* **9**, 2603 (2020).
49. Hu, B. *et al.* Neuronal and glial 3D chromatin architecture informs the cellular etiology of brain disorders. *Nat Commun* **12**, 3968 (2021).
50. Eastwood, S. L. *et al.* Decreased expression of mRNAs encoding non-NMDA glutamate receptors GluRl and GluR2 in medial temporal lobe neurons in schizophrenia. *Molecular Brain Research* **29**, 211–223 (1995).
51. Choi, K. H., Zepp, M. E., Higgs, B. W., Weickert, C. S. & Webster, M. J. Expression profiles of schizophrenia susceptibility genes during human prefrontal cortical development. *J Psychiatry Neurosci* **34**, 450–458 (2009).
52. Benesh, J. L., Mueller, T. M. & Meador-Woodruff, J. H. AMPA receptor subunit localization in schizophrenia anterior cingulate cortex. *Schizophrenia Research* **249**, 16–24 (2022).
53. Martin-Vilchez, S. *et al.* RhoGTPase Regulators Orchestrate Distinct Stages of Synaptic Development. *PLoS ONE* **12**, e0170464 (2017).
54. Bloch-Gallego, E. & Anderson, D. I. Key role of Rho GTPases in motor disorders associated with neurodevelopmental pathologies. *Mol Psychiatry* **28**, 118–126 (2023).
55. Townsend, L. N. *et al.* Cdk12 maintains the integrity of adult axons by suppressing actin remodeling. *Cell Death Discov.* **9**, 348 (2023).

56. Gardiner, E. J. *et al.* Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. *Journal of Psychiatric Research* **47**, 425–437 (2013).
57. Shoval, G., Bar-Shira, O., Zalsman, G., John Mann, J. & Chechik, G. Transitions in the transcriptome of the serotonergic and dopaminergic systems in the human brain during adolescence. *European Neuropsychopharmacology* **24**, 1123–1132 (2014).
58. Jamu, I. M. & Okamoto, H. Recent advances in understanding adverse effects associated with drugs targeting the serotonin receptor, 5-HT GPCR. *Front. Glob. Womens Health* **3**, 1012463 (2022).
59. Nishioka, M. *et al.* Comprehensive DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia. *J Hum Genet* **58**, 91–97 (2013).
60. Tao, R. *et al.* Cannabinoid receptor CNR1 expression and DNA methylation in human prefrontal cortex, hippocampus and caudate in brain development and schizophrenia. *Transl Psychiatry* **10**, 158 (2020).
61. D'Addario, C. *et al.* A preliminary study of endocannabinoid system regulation in psychosis: Distinct alterations of CNR1 promoter DNA methylation in patients with schizophrenia. *Schizophrenia Research* **188**, 132–140 (2017).
62. Urigüen, L. *et al.* Immunodensity and mRNA expression of A2A adenosine, D2 dopamine, and CB1 cannabinoid receptors in postmortem frontal cortex of subjects with schizophrenia: effect of antipsychotic treatment. *Psychopharmacology* **206**, 313–324 (2009).
63. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat Genet* **52**, 254–263 (2020).
64. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**, 1061 (2020).
65. Mouri, K. *et al.* Whole-genome functional characterization of RE1 silencers using a modified massively parallel reporter assay. *Cell Genomics* **3**, 100234 (2023).
66. Huang, D. & Ovcharenko, I. Enhancer–silencer transitions in the human genome. *Genome Res.* **32**, 437–448 (2022).
67. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* (2013) doi:10.1016/j.cell.2013.11.033.
68. Warren, T. L., Lambert, J. T. & Nord, A. S. AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *JoVE* e62650 (2022) doi:10.3791/62650.
69. Lagunas, T. *et al.* A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of non-coding elements in vivo. *Commun Biol* **6**, 1151 (2023).
70. Brown, A. R. *et al.* An *in vivo* massively parallel platform for deciphering tissue-specific regulatory function. Preprint at https://doi.org/10.1101/2022.11.23.517755 (2022).
71. Farrow, S. L. *et al.* Identification of 27 allele-specific regulatory variants in Parkinson's disease using a massively parallel reporter assay. *npj Parkinsons Dis.* **10**, 44 (2024).
72. Deng, C. *et al.* Massively parallel characterization of regulatory elements in the developing human cortex. *Science* **384**, eadh0559 (2024).
73. Van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* **51**, 1160–1169 (2019).

74. Lam, M. *et al.* Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *The American Journal of Human Genetics* **105**, 334–350 (2019).
75. Shi, Y. *et al.* Mouse and human share conserved transcriptional programs for interneuron development. *Science* **374**, eabj6641 (2021).
76. Diehl, A. G. & Boyle, A. P. Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse. *Nucleic Acids Research* **46**, 1878–1894 (2018).
77. Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, e04837 (2015).
78. Chini, M. & Hanganu-Opatz, I. L. Prefrontal Cortex Development in Health and Disease: Lessons from Rodents and Humans. *Trends in Neurosciences* **44**, 227–240 (2021).
79. Zeiss, C. J. Comparative Milestones in Rodent and Human Postnatal Central Nervous System Development. *Toxicol Pathol* **49**, 1368–1373 (2021).
80. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).
81. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research* **51**, D1188–D1195 (2023).
82. The ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
83. Farrell, C. M. *et al.* RefSeq Functional Elements as experimentally assayed nongenic reference standards and functional interactions in human and mouse. *Genome Res.* **32**, 175–188 (2022).
84. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**, 418–420 (2000).
85. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
86. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
87. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
88. Loo, M., P. J. ,van,der. The stringdist Package for Approximate String Matching. *The R Journal* **6**, 111 (2014).
89. Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* **10**, 707 (1966).
90. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE* **16**, e0257521 (2021).
91. 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).
92. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676–682 (2012).
93. Cheng, C., Fei, Z. & Xiao, P. Methods to improve the accuracy of next-generation sequencing. *Front. Bioeng. Biotechnol.* **11**, 982111 (2023).

94. Bailey, T. L. & Grant, C. E. SEA: Simple Enrichment Analysis of motifs. Preprint at https://doi.org/10.1101/2021.08.23.457422 (2021).
95. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).
96. Kulakovskiy, I. V. *et al.* HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res* **44**, D116–D125 (2016).
97. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**, D174–D182 (2024).
98. Simon Gert Coetzee [Aut, C. D. J. H. [Aut. motifbreakR. Bioconductor https://doi.org/10.18129/B9.BIOC.MOTIFBREAKR (2017).
99. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165–D173 (2022).
100. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research* **41**, D996–D1008 (2012).
101. Theiler, K. *The House Mouse: Atlas of Embryonic Development*. (Springer Berlin Heidelberg, Berlin, Heidelberg, 1989). doi:10.1007/978-3-642-88418-4.
102. Baldarelli, R. M. *et al.* The mouse Gene Expression Database (GXD): 2021 update. *Nucleic Acids Research* **49**, D924–D931 (2021).

Supplementary files

Fig. S1: CREs show moderate-to-strong correlations across mouse replicates. (A,B) Proportions of read counts per CRE in the (A) DNA and (B) RNA are strong across mice, with Spearman's rho ranging from 0.77-0.97. (C) Activity ratios are more moderately correlated and show comparable correlation coefficients to those reported in previous MPRAs[30].

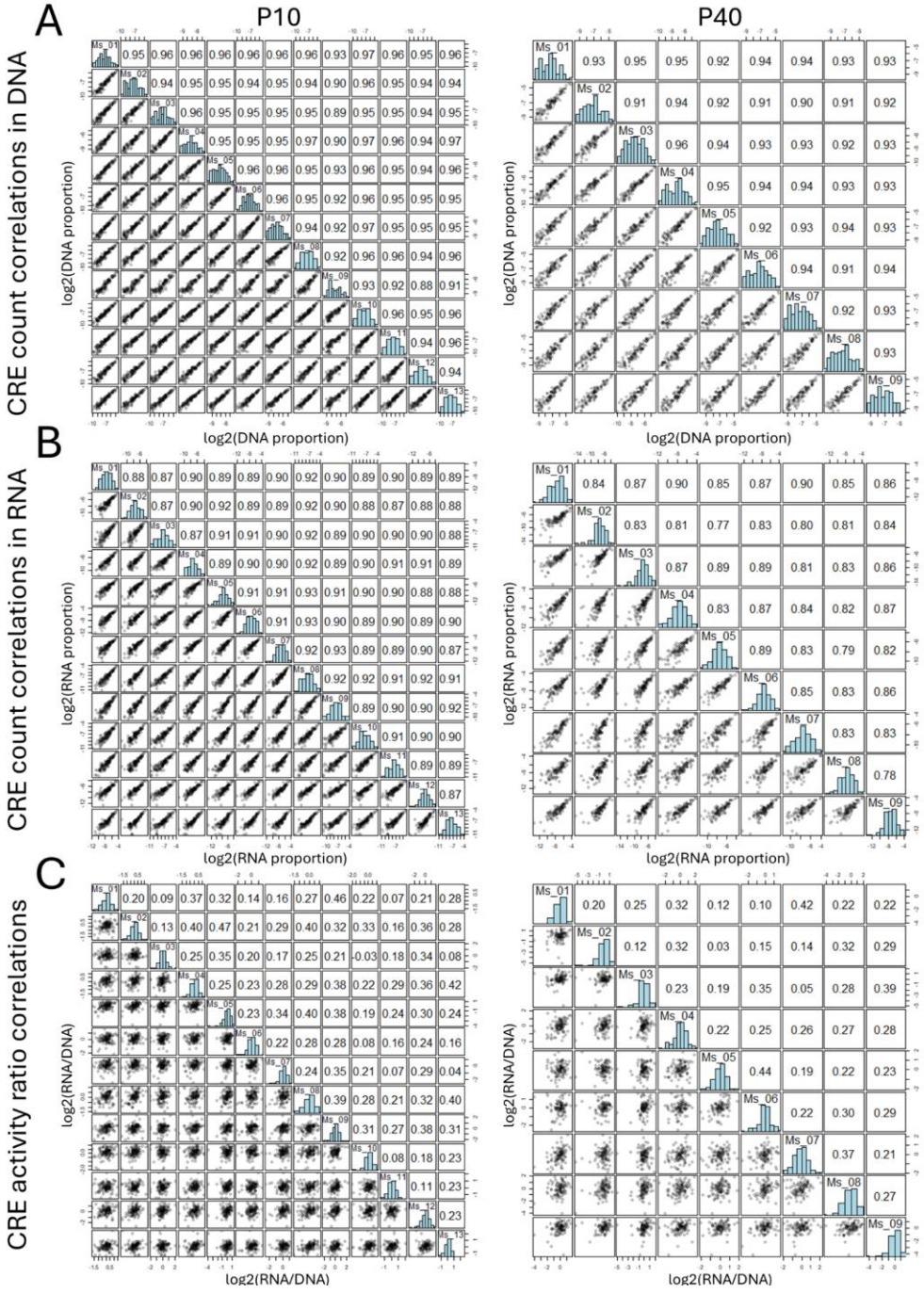


Fig. S2: FDR-significant CREs which are not highlighted in the main text. Coordinates and SNPs can be found in Tables 1 and S1.

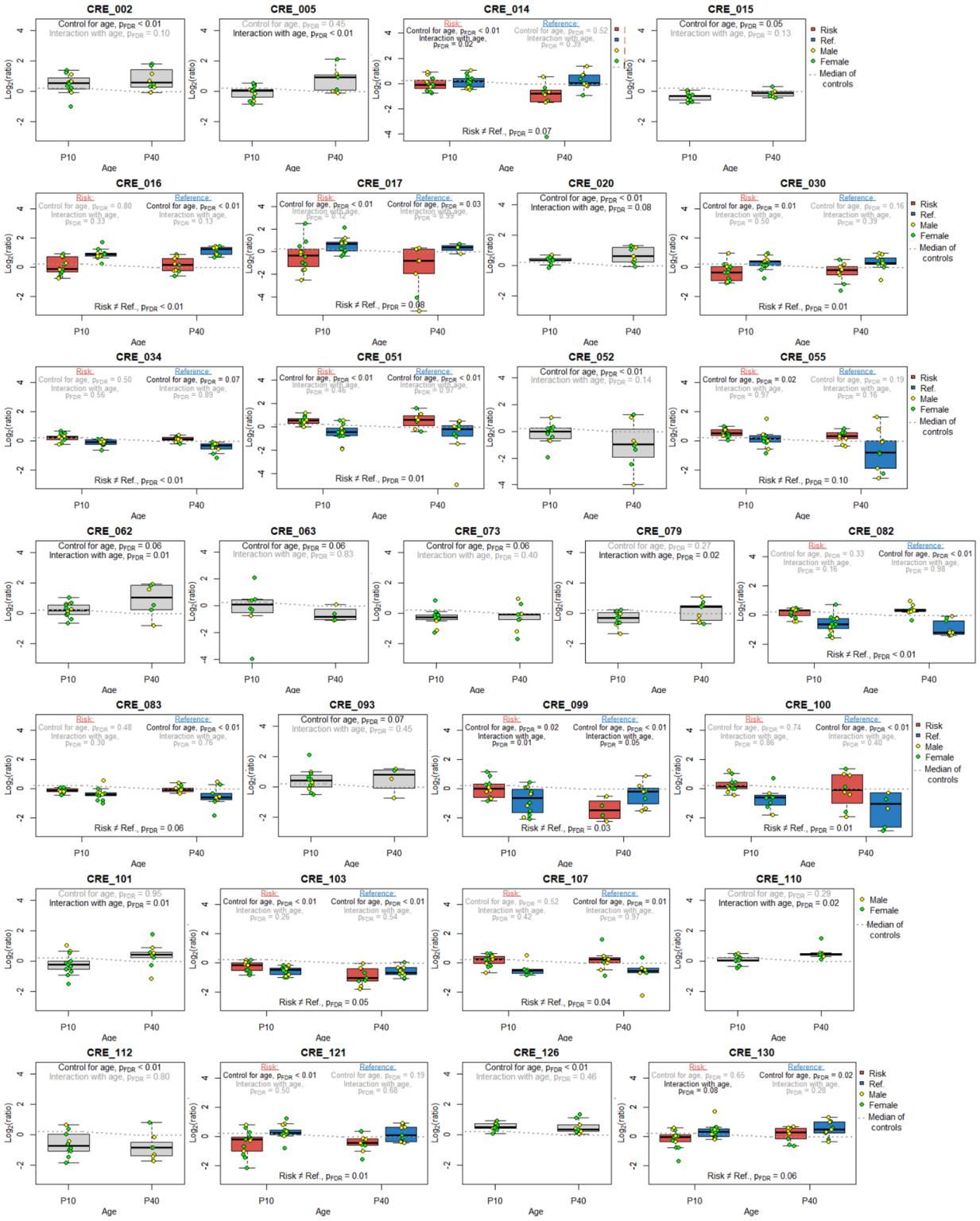


Fig. S3: Activity for individual barcodes representing CREs highlighted in the main text. (A) CREs in Figure 4. (B) CREs in Figure 5. (C) CREs in Figure 6. Dots represent individual animals and colors represent individual barcodes. CREs are indicated by blue arrows pointing upward if they were identified as enhancers. CREs identified as silencers are indicated by red arrows pointing downward. CREs which showed adolescent-specific upregulation are indicated by purple diagonal arrows. Barcodes showed evidence of linear, though somewhat noisy, relationships between $\log_2$(RNA) and $\log_2$(DNA). Beta values for $\log_2$(RNA) versus $\log_2$(DNA) were higher for enhancers than for silencers. This indicates increased RNA returned per amount of MPRA DNA initially tested. Similarly, CREs which showed upregulation from P10 to P40 show higher beta values at P40. All CREs are represented by at least five barcodes in each animal and no barcode appears to dominate the activity profile for any CRE. Number of barcodes per CRE was controlled for in calculating $\log_2$(RNA/DNA) values used in the text.

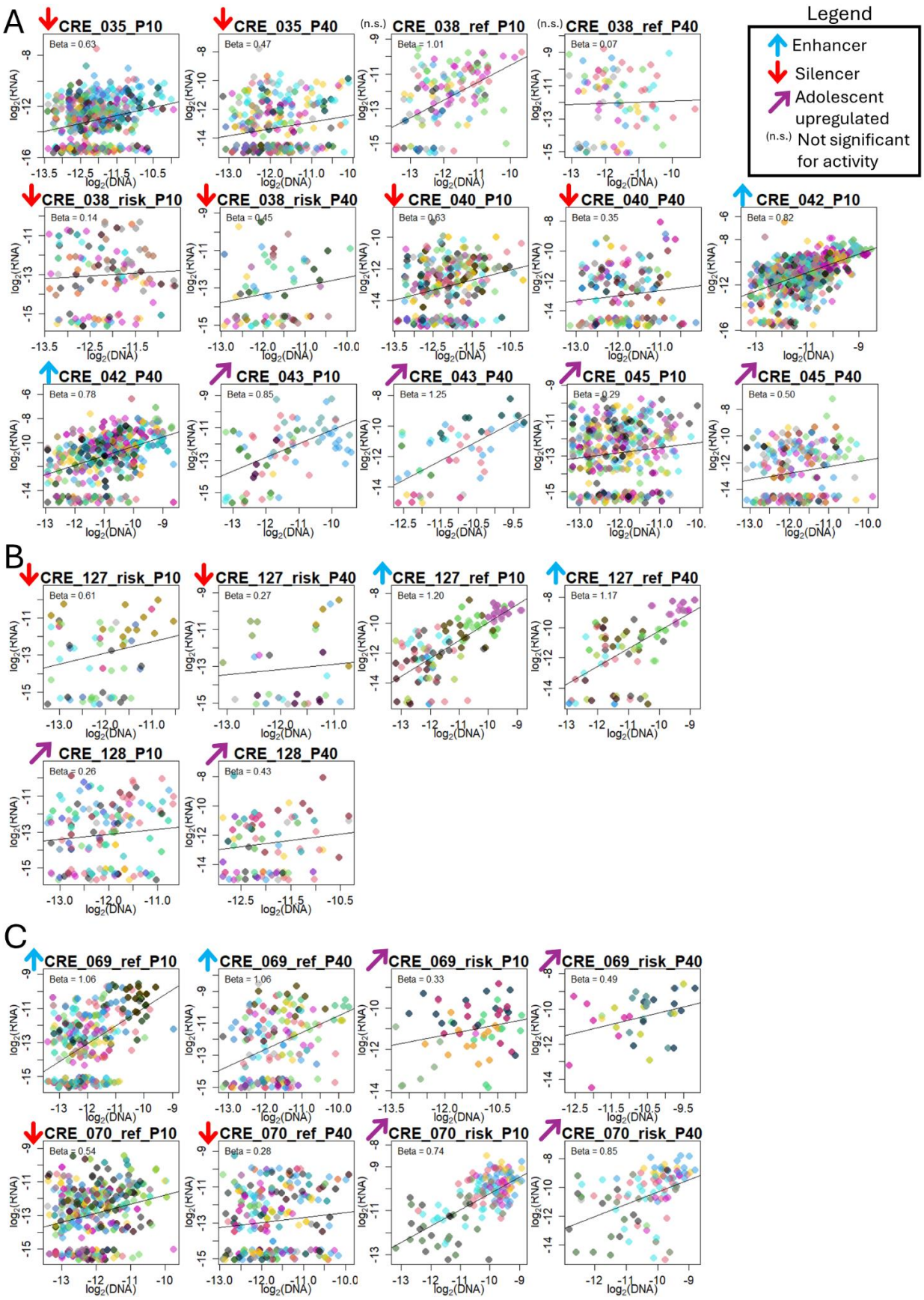

A
CRE_035_P10
Beta = 0.63
CRE_035_P40
Beta = 0.47
(n.s.) CRE_038_ref_P10
Beta = 1.01
(n.s.) CRE_038_ref_P40
Beta = 0.07
CRE_038_risk_P10
Beta = 0.14
CRE_038_risk_P40
Beta = 0.45
CRE_040_P10
Beta = 0.63
CRE_040_P40
Beta = 0.35
CRE_042_P10
Beta = 0.82
CRE_042_P40
Beta = 0.78
CRE_043_P10
Beta = 0.85
CRE_043_P40
Beta = 1.25
CRE_045_P10
Beta = 0.29
CRE_045_P40
Beta = 0.50
Legend
Enhancer
Silencer
Adolescent upregulated
(n.s.) Not significant for activity
B
CRE_127_risk_P10
Beta = 0.61
CRE_127_risk_P40
Beta = 0.27
CRE_127_ref_P10
Beta = 1.20
CRE_127_ref_P40
Beta = 1.17
CRE_128_P10
Beta = 0.26
CRE_128_P40
Beta = 0.43
C
CRE_069_ref_P10
Beta = 1.06
CRE_069_ref_P40
Beta = 1.06
CRE_069_risk_P10
Beta = 0.33
CRE_069_risk_P40
Beta = 0.49
CRE_070_ref_P10
Beta = 0.54
CRE_070_ref_P40
Beta = 0.28
CRE_070_risk_P10
Beta = 0.74
CRE_070_risk_P40
Beta = 0.85
log2(RNA)
log2(DNA)

Table S1: Results from all MPRA tests.

CREs	Coordinates (hg38)	SNP(s)	Risk allele(s)	Ref. allele(s)	Alleles tested?	pFDR for allelic differences		CRE activity											
								Without alleles				Risk allele				Reference allele			
								Controlling for age		Interaction with age		Controlling for age		Interaction with age		Controlling for age		Interaction with age	
						Ctrl. for age	Interaction with age	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p	Beta	FDR-p
CRE_002	chr1: 27148853-27149295	rs11247612	C	T	FALSE	NA	NA	0.55	0.00	0.02	0.10	NA	NA	NA	NA	NA	NA	NA	NA
CRE_003	chr1: 153802316-153802739	rs4077168	A	T	FALSE	NA	NA	0.16	0.42	0.01	0.68	NA	NA	NA	NA	NA	NA	NA	NA
CRE_005	chr1: 205079950-205080424	rs9651228	G	A	FALSE	NA	NA	0.15	0.45	0.04	0.00	NA	NA	NA	NA	NA	NA	NA	NA
CRE_007	chr1: 239056507-239056920	rs12119613	C	T	FALSE	NA	NA	-0.01	0.97	-0.02	0.14	NA	NA	NA	NA	NA	NA	NA	NA
CRE_008	chr2: 48315648-48316116	rs13383364	C	G	FALSE	NA	NA	0.11	0.61	0.00	0.76	NA	NA	NA	NA	NA	NA	NA	NA
CRE_013	chr2: 48337534-48337990	rs57729372	T	A	TRUE	0.18	0.11	-0.25	0.14	0.00	1.00	NA	NA	NA	NA	NA	NA	NA	NA
CRE_014	chr2: 48337966-48338436	rs59761994	G	A	TRUE	0.07	0.10	NA	NA	NA	NA	-0.54	0.00	-0.03	0.02	0.12	0.52	0.01	0.39
CRE_015	chr2: 48339556-48340049	rs6545034	C	T	TRUE	0.20	0.53	-0.32	0.05	0.02	0.13	NA	NA	NA	NA	NA	NA	NA	NA
CRE_016	chr2: 48341409-48341886	rs62137002	G	A	TRUE	0.00	0.75	NA	NA	NA	NA	0.05	0.80	0.01	0.33	0.94	0.00	0.02	0.13
CRE_017	chr2: 48354574-48355005	rs11685861	T	C	TRUE	0.08	0.76	NA	NA	NA	NA	-0.83	0.00	-0.02	0.12	0.42	0.03	0.00	0.99
CRE_018	chr2: 48356671-48357006	rs60527991	C	T	TRUE	0.54	0.26	-0.03	0.89	0.01	0.33	NA	NA	NA	NA	NA	NA	NA	NA
CRE_019	chr2: 48356983-48357388	rs11690690	G	A	TRUE	0.99	0.05	NA	NA	NA	NA	-0.13	0.47	-0.01	0.30	-0.07	0.71	0.01	0.40
CRE_020	chr2: 48358463-48358924	rs7563625	A	C	FALSE	NA	NA	0.45	0.00	0.02	0.08	NA	NA	NA	NA	NA	NA	NA	NA
CRE_023	chr2: 134711875-134712374	rs6747870	A	G	FALSE	NA	NA	0.11	0.58	0.00	0.74	NA	NA	NA	NA	NA	NA	NA	NA
CRE_024	chr2: 134725614-134726038	rs948583	G	A	FALSE	NA	NA	0.12	0.53	0.01	0.21	NA	NA	NA	NA	NA	NA	NA	NA
CRE_027	chr2: 171778232-171778707	rs35391289	A	C	TRUE	0.80	0.26	-0.31	0.12	0.02	0.33	NA	NA	NA	NA	NA	NA	NA	NA
CRE_029	chr4: 87523500-87523902	rs62315965	A	G	FALSE	NA	NA	0.10	0.58	0.01	0.41	NA	NA	NA	NA	NA	NA	NA	NA
CRE_030	chr4: 87523885-87524351	rs4693832	A	G	TRUE	0.01	0.90	NA	NA	NA	NA	-0.39	0.01	0.01	0.50	0.24	0.16	0.01	0.39
CRE_032	chr4: 90078405-90078807	rs6839192	T	C	TRUE	0.74	0.99	-0.18	0.33	0.00	0.93	NA	NA	NA	NA	NA	NA	NA	NA
CRE_034	chr5: 64401976-64402366	rs898519	G	A	TRUE	0.00	0.26	NA	NA	NA	NA	0.12	0.50	0.01	0.56	-0.30	0.07	0.00	0.89
CRE_035	chr5: 153924782-153925243	rs13170510	G	A	TRUE	0.81	0.26	-0.48	0.00	0.00	0.83	NA	NA	NA	NA	NA	NA	NA	NA
CRE_038	chr5: 153951427-153951885	rs35428024	G	A	TRUE	0.04	0.58	NA	NA	NA	NA	-0.52	0.00	-0.01	0.71	0.20	0.28	0.01	0.40

CRE_039	chr5:153955422-153955838	rs1025260	A	T	FALSE	NA	NA	0.19	0.32	0.00	0.86	NA	NA	NA	NA	NA	NA	NA	NA
CRE_040	chr5:153963255-153963754	rs58798017, rs66607728	T, C	C, G	FALSE	NA	NA	-0.30	0.08	0.02	0.19	NA	NA	NA	NA	NA	NA	NA	NA
CRE_042	chr5:153987019-153987412	rs425263	A	G	FALSE	NA	NA	0.37	0.02	0.01	0.35	NA	NA	NA	NA	NA	NA	NA	NA
CRE_043	chr5:153988354-153988751	rs453034	G	T	FALSE	NA	NA	-0.21	0.23	0.03	0.00	NA	NA	NA	NA	NA	NA	NA	NA
CRE_044	chr5:154002416-154002784	rs2560060	C	T	TRUE	0.26	0.90	0.06	0.74	0.01	0.40	NA	NA	NA	NA	NA	NA	NA	NA
CRE_045	chr5:154009899-154010396	rs12651907	A	G	TRUE	0.50	0.26	-0.01	0.97	0.02	0.06	NA	NA	NA	NA	NA	NA	NA	NA
CRE_046	chr5:154014138-154014494	rs690027	A	G	TRUE	0.91	0.54	-0.20	0.25	0.00	0.80	NA	NA	NA	NA	NA	NA	NA	NA
CRE_047	chr5:161905636-161906104	rs6865219	A	G	TRUE	0.72	0.72	-0.15	0.41	0.00	0.95	NA	NA	NA	NA	NA	NA	NA	NA
CRE_049	chr6:27805867-27806263	rs200481	A	T	TRUE	0.77	0.90	0.08	0.68	0.00	0.95	NA	NA	NA	NA	NA	NA	NA	NA
CRE_050	chr6:27806959-27807437	rs200483	A	G	FALSE	NA	NA	-0.11	0.54	0.01	0.57	NA	NA	NA	NA	NA	NA	NA	NA
CRE_051	chr6:31040643-31041102	rs2517549	A	C	TRUE	0.01	0.81	NA	NA	NA	NA	0.51	0.00	0.01	0.46	-0.64	0.00	0.00	0.97
CRE_052	chr6:31041230-31041704	rs2523901	C	T	FALSE	NA	NA	-0.56	0.00	-0.02	0.14	NA	NA	NA	NA	NA	NA	NA	NA
CRE_055	chr6:31043055-31043554	rs2844647, rs2248386	A, T	G, C	TRUE	0.10	0.48	NA	NA	NA	NA	0.38	0.02	0.00	0.97	-0.25	0.19	-0.02	0.16
CRE_057	chr6:31046898-31047395	rs2517536	A	T	TRUE	0.77	0.37	0.22	0.21	0.01	0.40	NA	NA	NA	NA	NA	NA	NA	NA
CRE_059	chr6:31048149-31048586	rs2844644	A	G	TRUE	0.04	NA	NA	NA	NA	NA	0.20	0.27	0.00	0.94	NA	NA	NA	NA
CRE_061	chr6:31049042-31049355	rs2251830	A	C	FALSE	NA	NA	0.25	0.15	0.00	0.74	NA	NA	NA	NA	NA	NA	NA	NA
CRE_062	chr6:31049481-31049854	rs2523866, rs2844643	C, A	T, G	FALSE	NA	NA	0.34	0.06	0.03	0.01	NA	NA	NA	NA	NA	NA	NA	NA
CRE_063	chr6:31050404-31050890	rs2523865, rs2523864	C, T	A, C	FALSE	NA	NA	-0.44	0.06	0.00	0.83	NA	NA	NA	NA	NA	NA	NA	NA
CRE_069	chr6:87898106-87898580	rs9450822	G	A	TRUE	0.04	0.19	NA	NA	NA	NA	0.52	0.00	0.01	0.37	0.06	0.75	0.03	0.00
CRE_070	chr6:87900187-87900655	rs4235834	G	T	TRUE	0.00	0.05	NA	NA	NA	NA	0.24	0.16	0.02	0.06	-0.36	0.02	0.00	0.97
CRE_071	chr6:87901909-87902384	rs6918951, rs6903270	T, C	C, T	FALSE	NA	NA	0.07	0.74	0.01	0.23	NA	NA	NA	NA	NA	NA	NA	NA
CRE_073	chr6:100605037-100605531	rs240115	C	T	TRUE	0.78	0.23	-0.32	0.06	0.01	0.40	NA	NA	NA	NA	NA	NA	NA	NA
CRE_074	chr6:100610788-100611286	rs240113	A	G	FALSE	NA	NA	0.01	0.98	0.00	0.98	NA	NA	NA	NA	NA	NA	NA	NA
CRE_075	chr6:100617672-100618171	rs240154	T	G	TRUE	0.13	0.40	-0.13	0.46	0.02	0.20	NA	NA	NA	NA	NA	NA	NA	NA
CRE_076	chr6:100630337-100630783	rs240147	A	G	TRUE	0.94	0.74	0.05	0.80	0.00	0.97	NA	NA	NA	NA	NA	NA	NA	NA
CRE_077	chr6:100630708-100631099	rs240146	A	G	TRUE	0.98	0.72	0.04	0.86	0.01	0.21	NA	NA	NA	NA	NA	NA	NA	NA

CRE_078	chr6: 100631483-100631923	rs240145	C	A	TRUE	0.18	0.74	-0.07	0.73	0.00	0.84	NA	NA	NA	NA	NA	NA	NA	NA
CRE_079	chr6: 100632627-100633112	rs240143	A	G	TRUE	0.21	0.84	-0.20	0.27	0.03	0.02	NA	NA	NA	NA	NA	NA	NA	NA
CRE_082	chr6: 100640959-100641447	rs239246	A	G	TRUE	0.00	0.23	NA	NA	NA	NA	0.17	0.33	0.02	0.16	-0.79	0.00	0.00	0.98
CRE_083	chr6: 100642200-100642573	rs239245	A	C	TRUE	0.06	0.54	NA	NA	NA	NA	-0.13	0.48	0.01	0.30	-0.49	0.00	0.00	0.76
CRE_093	chr7: 127475956-127476260	rs6957647	A	G	FALSE	NA	NA	0.36	0.07	0.01	0.45	NA	NA	NA	NA	NA	NA	NA	NA
CRE_094	chr7: 127485895-127486341	rs33957168	A	G	TRUE	0.81	0.23	-0.12	0.49	0.01	0.65	NA	NA	NA	NA	NA	NA	NA	NA
CRE_098	chr7: 127491411-127491856	rs17150921	T	A	TRUE	0.13	0.71	-0.14	0.44	0.01	0.56	NA	NA	NA	NA	NA	NA	NA	NA
CRE_099	chr7: 127493877-127494356	rs34976296	T	C	TRUE	0.72	0.03	NA	NA	NA	NA	-0.48	0.02	-0.04	0.01	-0.69	0.00	0.02	0.05
CRE_100	chr7: 127494107-127494580	rs13228307	A	G	TRUE	0.01	0.78	NA	NA	NA	NA	0.07	0.74	0.00	0.86	-1.01	0.00	-0.01	0.40
CRE_101	chr7: 127498321-127498799	rs1419401	G	T	FALSE	NA	NA	-0.02	0.95	0.03	0.01	NA	NA	NA	NA	NA	NA	NA	NA
CRE_103	chr7: 127501956-127502455	rs6975687, rs6979978	A, C	G, G	TRUE	0.72	0.05	NA	NA	NA	NA	-0.57	0.00	-0.01	0.26	-0.63	0.00	0.01	0.54
CRE_104	chr8: 22387204-22387673	rs11783730	T	C	FALSE	NA	NA	-0.08	0.69	0.02	0.14	NA	NA	NA	NA	NA	NA	NA	NA
CRE_107	chr8: 22411059-22411542	rs11780055	A	T	TRUE	0.04	0.74	NA	NA	NA	NA	0.12	0.52	0.01	0.42	-0.58	0.01	0.00	0.97
CRE_110	chr8: 22431845-22432148	rs4872497	C	G	TRUE	0.88	0.85	0.19	0.29	0.03	0.02	NA	NA	NA	NA	NA	NA	NA	NA
CRE_112	chr8: 27400371-27400863	rs2163174	A	G	FALSE	NA	NA	-0.74	0.00	0.00	0.80	NA	NA	NA	NA	NA	NA	NA	NA
CRE_113	chr8: 42143499-42143887	rs60179949	C	T	FALSE	NA	NA	0.10	0.58	0.01	0.49	NA	NA	NA	NA	NA	NA	NA	NA
CRE_115	chr8: 42149019-42149481	rs4615558	A	G	TRUE	0.26	0.13	0.05	0.80	0.02	0.13	NA	NA	NA	NA	NA	NA	NA	NA
CRE_121	chr10: 48831865-48832352	rs7908454	G	A	TRUE	0.01	0.90	NA	NA	NA	NA	-0.51	0.00	0.01	0.50	0.23	0.19	0.01	0.68
CRE_124	chr16: 49005944-49006389	rs715017	G	C	TRUE	0.74	0.22	0.02	0.94	0.02	0.13	NA	NA	NA	NA	NA	NA	NA	NA
CRE_125	chr16: 49007570-49008046	rs1477099	C	T	TRUE	0.90	0.34	-0.10	0.57	0.01	0.40	NA	NA	NA	NA	NA	NA	NA	NA
CRE_126	chr17: 21379811-21380270	rs1317349	T	G	TRUE	0.63	0.37	0.48	0.00	0.01	0.46	NA	NA	NA	NA	NA	NA	NA	NA
CRE_127	chr17: 39276765-39277245	rs59585619	G	C	TRUE	0.02	0.90	NA	NA	NA	NA	-0.35	0.08	0.01	0.33	0.43	0.01	0.01	0.38
CRE_128	chr17: 39277370-39277868	rs56058902	T	C	TRUE	0.26	0.86	-0.24	0.20	0.02	0.08	NA	NA	NA	NA	NA	NA	NA	NA
CRE_129	chr18: 53420377-53420875	rs11082997	G	A	TRUE	0.26	0.94	-0.22	0.20	0.01	0.46	NA	NA	NA	NA	NA	NA	NA	NA
CRE_130	chr18: 53429778-53430268	rs2879527	C	T	TRUE	0.06	0.75	NA	NA	NA	NA	-0.09	0.65	0.02	0.08	0.37	0.02	0.01	0.28

Table S3: MPRA barcodes. Ns indicate wildcards which are replaced with all four nucleotides (A, C, G, T) during oligo synthesis.

Barcode ID	Barcode sequence
Barcode_01	NCCNAGTNCCTNTGNCTCNAGNCCG
Barcode_02	NCTNAGCNAGGNCTNTCGNTGNCTG
Barcode_03	NGCNAAGNCTCNTCNTCTNCGNTGC
Barcode_04	NTANCGCNACTNGCNTCGNTGNCTG
Barcode_05	NTTNGCCNTCGNTCNGTCNGCNACT
Barcode_06	NTCNGGCNGTCNAGNCTGNCTNCTA
Barcode_07	NCGNAACNGTCNAGNCCANTGNCCT
Barcode_08	NCGNCTCNTCTNCANTTGNCGNTCG
Barcode_09	NAGNTCCNTCTNTGNCTCNTGNCCT
Barcode_10	NCGNAACNGCCNTANTCTNCGNTAC
Barcode_11	NCGNATCNGCTNCANCGTNACNATC
Barcode_12	NGCNTCGNATTNGCNTCTNCANCGT
Barcode_13	NCGNAGCNGAANTGNCTCNATNCGT
Barcode_14	NCGNACTNGTCNTCNGTCNGCNATT
Barcode_15	NCGNATANCGCNATNCGTNTCNGTC
Barcode_16	NCGNCTANTTCNGTNCTCNTCNGCT
Barcode_17	NCGNATCNGACNTGNCTGNCTNCAA
Barcode_18	NCGNACCNGCTNATNGTCNATNCGT
Barcode_19	NGCNTCTNCGCNTCNGTANTCNTCT
Barcode_20	NGCNTCTNCCTNGCNTCTNTGNTCT
Barcode_21	NCGNAACNGTCNTGNTCCNATNGTC
Barcode_22	NAANGCTNCTCNATNGTCNTGNCTG
Barcode_23	NCGNATANCGTNGCNTCANTGNCTT
Barcode_24	NCGNATANGTCNAGNCTCNTGNTCT
Barcode_25	NCGNACTNGCTNCANTGCNTCNATA

Table S4: Validated negative controls

Neg. control	Genome assembly	Coordinates
Hm_neg_01	hg38	chr12:2235351-2236252
Hm_neg_02	hg38	chr12:2246176-2247090
Hm_neg_03	hg38	chr12:2274375-2275292
Hm_neg_04	hg38	chr12:2277430-2278308
Hm_neg_05	hg38	chr12:2285152-2286076
Hm_neg_06	hg38	chr12:2292626-2293505
Hm_neg_07	hg38	chr12:2298364-2299275
Hm_neg_08	hg38	chr12:2411046-2411965
Hm_neg_09	hg38	chr12:2413761-2414661
Ms_neg_01	mm10	chr1:87454318-87455211
Ms_neg_02	mm10	chr4:99401841-99402751
Ms_neg_03	mm10	chr5:3090657-3091561

Ms_neg_04	mm10	chr7:25891471-25892364
Ms_neg_05	mm10	chr7:10199094-10199972
Ms_neg_06	mm10	chr11:41121892-41122770
Ms_neg_07	mm10	chr11:43088595-43089473
Ms_neg_08	mm10	chr13:11117678-11118588
Ms_neg_09	mm10	chr13:75244237-75245134
Ms_neg_10	mm10	chr14:13138844-13139737
Ms_neg_11	mm10	chr14:15078304-15079201
Ms_neg_12	mm10	chr16:50973438-50974336
Ms_neg_13	mm10	chr17:61585003-61585910

Chapter 3:
Identification of convergent pathways from common variants toward psychotic disorder endophenotypes

a. Preface

In the following study, I moved from testing the direct biochemical impacts of common psychiatric risk variants (Chapters 1 & 2) to testing their impact in aggregate on hypothesized convergent pathways toward disease. I worked with my co-first author on the resulting publication, Justin Tubbs, to adapt the traditional polygenic risk score (PGS) into the more biologically fine-grained pathway polygenic risk score (pPGS). In traditional PGS, an individual's risk variants for disease from across the genome are assigned specific weights corresponding to the strength of their association with disease, and subsequently summed to give a single metric of genomic disease risk for a given individual[1]. In pPGS, the genome is first subdivided into a gene set composed of genes relevant to a specific biological pathway of interest. Following this, the PGS is calculated as usual, but only using variants that are contained in or thought to affect genes within the identified gene set[2]. This results in the calculation of polygenic risk for disease that is hypothesized to affect a specific biological pathway, allowing for the direct testing of associations between specific pathways and disease symptoms in human subjects.

Using this technique, we studied PGS for schizophrenia (SZ) and SZ pPGS defined on different neurotransmitter systems for 205 demographically diverse human subjects with a psychotic disorder and 115 demographically diverse healthy controls. Subjects underwent intensive clinical, psychological, and neurological phenotyping at the UC Davis Medical Center in the lab of our collaborator, Dr. Cameron Carter. Subjects were genotyped on an Illumina SNP array using DNA extracted from whole blood. Using the most recent Psychiatric Genomics Consortium multi-ancestry GWAS for SZ[3], PGS was calculated for each subject, as well as

pPGS for variants thought to affect either glutamatergic, GABAergic, dopaminergic, or serotonergic synapses. While genome-wide PGS was strongly associated with overall disease risk, the pPGSs showed stronger relationships to particular disease endophenotypes and clinical outcomes. This is consistent with the hypothesis that SZ and other psychotic disorders may be biologically stratifiable into distinct etiological pathways. Our study shows the utility of pPGS for directly testing hypotheses of associations between specific biological pathways and psychiatric symptoms. We also provide evidence for differential treatment outcomes associated with genomic risk in different pathways. Treatment of disease etiologies as biologically stratifiable may lead to identification of new treatment targets and improved efficacy of novel therapeutics during clinical testing.

This study, for which I am co-first author, was conducted in collaboration with the labs of Dr. Cameron Carter at the UC Davis Medical Center and Dr. Pak Sham at the University of Hong Kong. It was published in *Molecular Psychiatry* on March 15, 2024[4]. I extracted and cleaned DNA from whole blood, which I then submitted for genotyping at the UC Davis Technologies Core. I conceived of testing for endophenotypes associated with pathway-specific genomic impacts and curated gene sets for each pPGS using several online gene ontology databases[5–7]. My co-first author, Justin Tubbs, calculated all PGS and pPGS as well as genomic ancestry principal components, based on 1000 Genomes[8] data. I corrected PGS and pPGS for ancestry-specific differences and conducted all subsequent analyses reported in the paper. I produced all tables, figures, and wrote the main text of the paper.

References

1. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 44 (2020).
2. Choi, S. W. *et al. The Power of Pathway-Based Polygenic Risk Scores*. https://www.researchsquare.com/article/rs-643696/v1 (2021) doi:10.21203/rs.3.rs-643696/v1.
3. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

4. Warren, T. L. *et al.* Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis. *Molecular Psychiatry* (2024) doi:10.1038/s41380-024-02457-0.
5. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27**, 29–34 (1999).
6. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289 (2009).
7. Joshi-Tope, G. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33**, D428–D432 (2004).
8. Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256–256 (2008).

b. Resulting Publication

DOI: https://doi.org/10.1038/s41380-024-02457-0

Mol Psychiatry, doi:10.3791/62650 (2022).

Title:

Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis

Author list:

Tracy L Warren *[,1,2,3], Justin D Tubbs *[,4,5,6,7], Tyler A Lesh[2], Mylena B Corona[1,2,3], Sarvenaz S Pakzad[2], Marina D Albuquerque[2], Praveena Singh[2], Vanessa Zarubin[2,8], Sarah J Morse[1,2,3,9], Pak Chung Sham [#,10,11,12], Cameron S Carter [#,13,14], Alex S Nord [#,15,16,17]

* Indicates co-first authors

[#] Indicates co-senior authors

Author affiliations:

[1] Department of Neurobiology, Physiology and Behavior, University of California, Davis, CA, USA.

[2] Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA, USA.

[3] Center for Neuroscience, University of California, Davis, CA, USA.

[4] Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR.

[5] Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.

[6] Department of Psychiatry, Harvard Medical School, Boston, MA, USA.

[7] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[8] Department of Psychology, Northwestern University, Evanston, IL, USA.

[9] Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA.

[10] Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR. pcsham@hku.hk.

[11] Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR. pcsham@hku.hk.

[12] State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR. pcsham@hku.hk.

[13] Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA, USA. cscarte1@hs.uci.edu.

[14] Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA. cscarte1@hs.uci.edu.

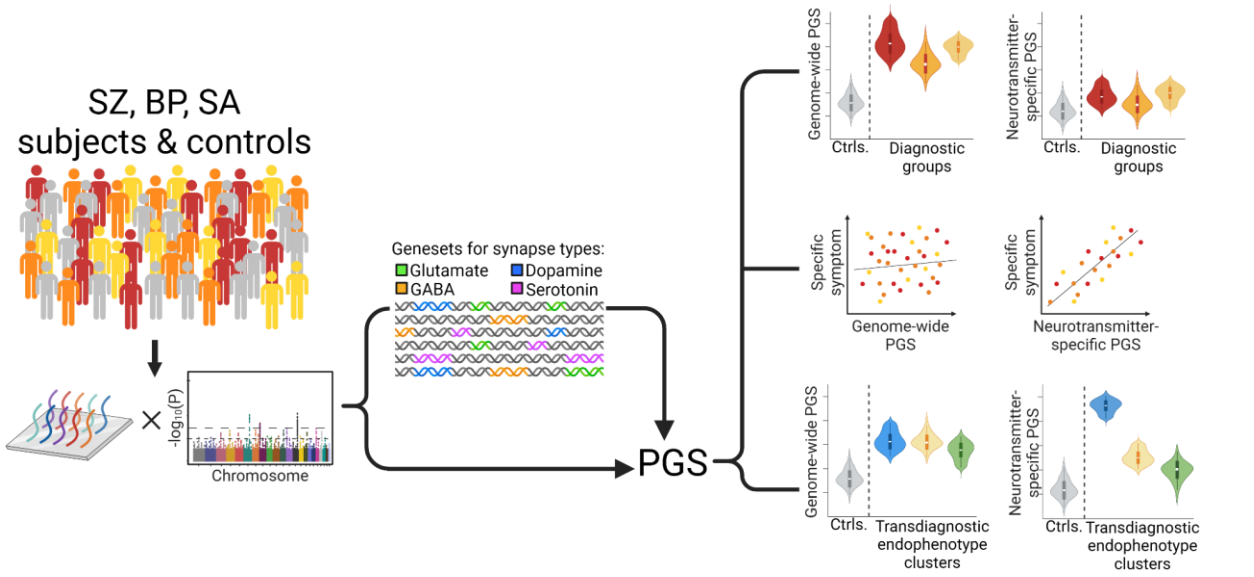
[15] Department of Neurobiology, Physiology and Behavior, University of California, Davis, CA, USA. asnord@ucdavis.edu.

[16] Department of Psychiatry and Behavioral Sciences, University of California, Davis, CA, USA. asnord@ucdavis.edu.

[17] Center for Neuroscience, University of California, Davis, CA, USA. asnord@ucdavis.edu.

Correspondence to Alex S. Nord (asnord@ucdavis.edu), Cameron S. Carter (cscarte1@hs.uci.edu), and Pak Chung Sham (pcsham@hku.hk).

Graphical abstract



Created with BioRender.com

Abstract

A primary goal of psychiatry is to better understand the pathways that link genetic risk to psychiatric symptoms. Here, we tested association of diagnosis and endophenotypes with overall and neurotransmitter pathway-specific polygenic risk in patients with early-stage psychosis. Subjects included 205 demographically diverse cases with a psychotic disorder who underwent comprehensive psychiatric and neurological phenotyping and 115 matched controls. Following genotyping, we calculated polygenic scores (PGSs) for schizophrenia (SZ) and bipolar disorder (BP) using Psychiatric Genomics Consortium GWAS summary statistics. To test if overall genetic risk can be partitioned into affected neurotransmitter pathways, we calculated pathway PGSs (pPGSs) for SZ risk affecting each of four major neurotransmitter systems: glutamate, GABA, dopamine, and serotonin. Psychosis subjects had elevated SZ PGS

versus controls; cases with SZ or BP diagnoses had stronger SZ or BP risk, respectively. There was no significant association within psychosis cases between individual symptom measures and overall PGS. However, neurotransmitter-specific pPGSs were moderately associated with specific endophenotypes; notably, glutamate was associated with SZ diagnosis and with deficits in cognitive control during task-based fMRI, while dopamine was associated with global functioning. Finally, unbiased endophenotype-driven clustering identified three diagnostically mixed case groups that separated on primary deficits of positive symptoms, negative symptoms, global functioning, and cognitive control. All clusters showed strong genome-wide risk. Cluster 2, characterized by deficits in cognitive control and negative symptoms, additionally showed specific risk concentrated in glutamatergic and GABAergic pathways. Due to the intensive characterization of our subjects, the present study was limited to a relatively small cohort. As such, results should be followed up with additional research at the population and mechanism level. Our study suggests pathway-based PGS analysis may be a powerful path forward to study genetic mechanisms driving psychiatric endophenotypes.

Introduction

Psychosis-spectrum disorders, including schizophrenia (SZ), schizoaffective disorder (SA), and bipolar disorder (BP), affect 0.5–2.3% of people worldwide [1,2,3,4,5]. Evidence suggests shared etiology for these disorders, supported by family studies and genome-wide association studies (GWASs) showing high interheritability and shared genomic risk [6,7,8]. Given symptom heterogeneity across the psychosis spectrum, a deeper understanding of the biology underlying specific symptoms may drive psychiatry towards improved patient outcomes using stratified medicine.

One approach toward revealing shared neurobiology is identifying transdiagnostic endophenotypes. Recently, the Bipolar & Schizophrenia Network for Intermediate Phenotypes

(B-SNIP) Consortium identified three biotypes of psychotic disorders using neuropsychiatric markers, agnostic to diagnoses [9], that were primarily characterized by differences in cognitive control and sensorimotor reactivity. As knowledge of biology underlying symptom heterogeneity expands, reclassifying psychosis patients using biologically grounded phenotypes may allow for more effective, targeted interventions.

Identification of causal and clinically informative genetic components of psychotic disorders has been aided by large-scale GWASs and estimation of overall genetic risk using polygenic scores (PGSs). While large cohorts were initially required to develop PGSs, once defined, research leveraging PGSs in cohorts with phenotyping that extends beyond diagnosis is revealing how genetic burden is associated with specific symptomatology. For example, recent efforts have found associations between SZ PGS and treatment response [10] as well as neurological and cognitive measures [11, 12]. Such uses of PGS enable study of genetic burden in small cohorts that have been characterized at a level not feasible at the scale required for GWASs focused on the discovery of new risk loci.

While PGS is a useful metric of genetic risk, it fails to assign burden into relevant biological pathways and elucidate mechanisms underlying symptoms. A novel approach to overcome this is pathway-specific PGS (pPGS), which partitions variants into those contained only in genes of a given function or pathway [13]. By assigning burden into pathways, it is possible to test how specific processes contribute to phenotype. This has been used to associate miR-137 pPGS with SZ risk [14] and neurological markers [15, 16]. Another recent implementation identified SZ and BP subjects with risk in pathways targetable by available pharmaceuticals, showing the utility of pPGS in targeted medicine [17]. Among pathways relevant to psychosis, glutamate [18,19,20,21,22,23], GABA [19,20,21,22,23,24], dopamine [23, 25,26,27,28], and serotonin [23, 28, 29] systems have been strongly associated with psychotic disorders and therapeutics.

There is a major opportunity to use pPGS to identify pathway contributions to psychosis symptoms. Here, we investigated the relationship of overall and neurotransmitter-associated PGS to psychotic disorder presentation at the diagnostic and endophenotypic level. We estimated overall and pathway-level PGS for a diverse cohort recruited from California following a first psychotic episode, and tested association between genetic burden, diagnosis, and clinical and neuroimaging endophenotypes. In a cohort of 205 subjects with psychotic disorders versus 115 controls, we found that overall and pathway PGSs were elevated in cases, pPGS was moderately associated with specific endophenotypes, and unbiased clustering on phenotypes showed associations with pPGS and treatment response. Our results demonstrate the power of pPGS to test the relationships between affected biological pathways and transdiagnostic phenotypes.

Methods

Study participants

Psychosis subjects were all outpatients within two years of their first psychotic episode. Subjects were selected from an ongoing psychosis research cohort, which includes 196 first-episode SZ-spectrum patients, 53 patients with first-episode BP with psychotic features, and 135 controls aged 12–38. After all quality control (QC), 118 SZ, 39 SA, 48 BP, and 115 controls were included for this study from an ongoing early psychosis research cohort. The study was approved by the University of California, Davis, Institutional Review Board and all subjects gave written consent and were paid for their participation.

Psychiatric and neuroimaging phenotyping

All participants were assessed using the Structured Clinical Interview for the DSM-IV-TR (SCID I/P) [30]. Clinical interviews were conducted by clinicians with masters or doctoral degrees

trained to high reliability (kappa > 0.70; range = 0.70–1.0). All patients were followed longitudinally and diagnoses were confirmed 6 months after ascertainment. Exclusion criteria for all groups included: Wechsler Abbreviated Scale of Intelligence (WASI) IQ score below 70, alcohol or drug dependence or abuse within 3 months before testing, positive urine toxicology screen for illicit drugs, prior head trauma worse than a Grade I concussion, or contraindication to MRI scanning. Control subjects were excluded for the following additional criteria: any lifetime diagnosis of an Axis I or Axis II disorder or any first-degree relatives with a psychotic disorder. Before testing, a detailed description of the study was provided and written informed consent obtained.

Subjects were evaluated on the Global Assessment of Functioning (GAF) [31], Global Social Functioning scale (GSF) [32], Global Role Functioning scale (GRF) [33], Young Mania Rating Scale (YMRS) [34], Scale for the Assessment of Positive Symptoms (SAPS) [35], Scale for the Assessment of Negative Symptoms (SANS) [35], and Brief Psychiatric Rating Scale (BPRS) [36]. Reality distortion, poverty symptoms, and disorganization scores were defined from the BPRS, SAPS, and SANS [37]. Treatment response was defined as >20% decrease in BPRS from baseline [38].

The GSF and GRF were measured at multiple time points; for each scale, we computed an average of the highest and lowest values measured for each subject during the past year and proceeded with these values for subsequent analyses. Finally, we removed item #8 (“Content”) from scores for the YMRS, as this question asks specifically about positive psychotic symptoms and can skew YMRS scores for subjects with schizophrenia. As such, scores on the YMRS used in these analyses more directly represent mania-specific symptoms. At baseline, all patients had BPRS scores ≥5 to offer sufficient resolution to detect a 20% improvement in score at follow-up. For treatment response calculation, BPRS was rescaled to a lowest score of zero (i.e. score of 24 = score of 0) [39].

Behavioral and neuroimaging methods are described previously [40]. Consequently, we present these methods in a condensed form. The AX-Continuous Performance Task (AX-CPT) [41] was performed during fMRI. In short, the task requires participants to respond to a series of cue and probe letters and correctly identify the target pair ("AX" trials) while correctly rejecting other pairs. The frequency manipulation of trial types creates a prepotent tendency to make a target response when the "X" probe letter is presented. Consequently when a non-A cue is presented and followed by an X (i.e. "BX" trials) the participant must engage proactive control to retain the goal, keep the incorrect cue in mind, and correctly reject the trial at the probe phase. Participants were excluded if performance did not meet a minimal threshold [42]. The primary behavioral measure used for this study is d'-context, which represents a contrast of AX hits versus BX false alarms.

Functional Blood Oxygenation Level Dependent (BOLD) data were acquired using a 1.5 T GE Signa and 3.0 T Siemens TimTrio. Two regions of interest were defined a priori and comprised CueB versus CueA contrast, reflecting high versus low cognitive control-related activity. Specifically, bilateral dorsolateral prefrontal cortex (DLPFC) and bilateral superior parietal cortex (SPC) were defined as 5 mm radius spheres based on coordinates from two independent datasets [43, 44]. All fMRI data were preprocessed using SPM8 (Wellcome Dept. of Imaging Neuroscience, London) and included slice timing correction, realignment, normalization to the Montreal Neurological Institute (MNI) template, and smoothing with an 8 mm FWHM Gaussian kernel. Individual runs were excluded when framewise displacement measures of movement exceeded 0.45 mm (calculated with https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLMotionOutliers) and whole subjects were excluded if more than half of their data exceeded this threshold. All trial types were modeled (CueA, CueB, AX, AY, BX, BY) and correct responses were used to create first-level images.

Based on the exclusion criteria mentioned above, five controls, nine SZ subjects, and two BP subjects were excluded due to excess motion. Two controls, five SZ subjects, and two BP subjects were excluded for poor behavioral performance. Finally, three controls and six SZ subjects were excluded for other reasons, including scanning artifacts, scanner failure, or button pad failure.

Genotyping, quality control, and imputation

Subjects underwent blood draws using PAXgene blood DNA tubes, which were subsequently stored at −80 °C until DNA extraction. Tubes were thawed at 37 °C for 15 min before extracting DNA using the Qiagen QIAamp DNA Blood Mini Kit. The protocol was followed with minor modifications: volumes of blood, protease, lysis buffer, and ethanol were tripled prior to binding DNA to the spin column; the final elution incubation was carried out at 50 °C for 5 min; and DNA was eluted in 100 µl of nuclease-free water. DNA was cleaned on the Zymo Research Genomic DNA Clean & Concentrator-10 kit with minor modifications: an extra 2-min spin in a clean collection tube was added after completing wash steps; samples were incubated at 50 °C for 5 min during elution; and DNA was eluted in 100 µl of nuclease-free water. After cleaning, samples were analyzed on a spectrophotometer and verified to have concentrations ≥50 ng/µl and 260/280 and 260/230 ratios ≥1.70.

DNA was genotyped using the Illumina (San Diego, California) Infinium PsychArray-24 Kit at the UC Davis DNA Technologies Core. Initial QC was performed using Illumina GenomeStudio following published guidance [45, 46]. Additional QC was applied using PLINK. First, variants with greater than 5% missingness were removed. We confirmed that no individuals had more than 1% of SNPs missing, nor did we observe any mismatch between genetic sex inferred by PLINK and subjects' self-reported sex. We removed variants with significantly different missingness rates between cases and controls ($p < 0.001$) or variants with significant deviation

from Hardy-Weinberg equilibrium ($p < 1e\text{-}6$). No severe heterozygosity outliers were observed. A small number of pairs of individuals appeared to be related; in these cases one subject from each pair was randomly excluded from subsequent analyses. Following imputation, the QC filters outlined above were performed again. Additionally, variants with imputation quality INFO scores <0.7 or empirical INFO scores <0.8 were removed.

PLINK (v1.9) was used to calculate genetic principal components (PCs) in the 1000 Genomes [47] Phase 3 dataset, which were projected onto our sample. Samples were submitted to the Michigan Imputation Server for genotype imputation, using the full 1000 Genomes [47] Version 3 dataset as the reference panel. Following all QC, 7,608,150 SNPs were available across 338 unrelated subjects. Following additional review, 14 subjects with schizophreniform disorder and 2 subjects with schizotypal disorder that had initially been included were excluded from subsequent analyses due to insufficient numbers, such that the final number of subjects was 320 as described in “Study Participants.”

We additionally checked for presence of known copy number variants (CNVs) of high penetrance for schizophrenia risk in our samples using iPsychCNV5 and PennCNV [48,49,50], using default parameters. CNVs called by both software were retained as consensus calls for further analysis. This resulted in a total of 333 CNVs across all subjects. We cross-referenced these results with genome-wide significant CNV loci associated with SZ [51] and filtered for CNVs in our sample that overlapped with at least 50% of one of these previously identified SZ risk CNV loci. One subject with SZ appeared to have a previously described SZ risk CNV: 15q11.2 deletion. As the role of deletion at 15q11.2 is currently of uncertain significance in schizophrenia risk [52, 53], we retained this individual for all analyses.

Polygenic score calculation

Using Psychiatric Genomics Consortium GWAS summary statistics for SZ (2021) [22] and BP (2022) [54], we employed PRS-CS [55] to calculate SZ and BP PGS, with the phi parameter set to 0.01 as recommended without a validation sample. PRS-CS applies shrinkage to optimize PGS prediction and may not be applicable when restricting PGS to genes from specific pathways. Therefore, we used the PRSet function from PRSice [56] to calculate pPGS for the four neurotransmitter pathways using SZ GWAS summary statistics [22], as our subjects were weighted toward SZ diagnosis and case-control status was best described by SZ PGS. As recommended by the authors of PRSet, in order to avoid eliminating SNPs in certain genes, no p-value filter was applied when calculating pPGS. In order to select the R^2 cutoff used for clumping, we tested five potential values (0.1, 0.3, 0.5, 0.7, 0.9) in the subset of European ancestry samples. All values returned a significant relationship with case-control status. For results reported here, we used a clumping R^2 of 0.7 as it showed the strongest association with case status. We report results using alternative clumping values in Fig. S2, which demonstrate stability of our findings across clumping R^2 values. Otherwise, default parameters of PRSet were retained, with the 1000 Genomes [47] European subset used as the LD reference panel.

To reduce bias from population stratification, we regressed calculated PGSs on the first 10 principal components of genetic variation in our subjects. We then assessed performance of regressed PGSs across different ancestries present in our study sample. Subject ancestries were estimated by assigning each subject to their nearest 1000 Genomes [47] population by minimizing Euclidean distance to population centroid in 20-dimensional PC space, as recommended by Privé et al. [57]. Regressed PGSs (hereafter called PGSs) showed substantial reduction in ancestry-specific stratification and appeared to replicate top-level associations when subsetted to just European/white subjects (Fig. S3).

pPGS pathways

Pathways include genes relevant in glutamate, GABA, dopamine, and serotonin. Genes were sourced from KEGG [58], REACTOME [59], and AmiGO [60] by searching for pathways and ontologies that include these neurotransmitters or variations on them (e.g. “glutamatergic”). The complete gene list is in Table S1.

Unbiased clustering

Endophenotypes were z-score normalized within cases. For cognitive control phenotypes, subjects were z-score normalized within cases with the same protocol/scanner type. Subjects were clustered on all endophenotypes. We used all subjects with complete endophenotype data ($n = 91$) to calculate the optimal number of clusters using the NbClust R package [61]. Estimates converged on three clusters. Subsequently, we used all subjects with greater than 50% of endophenotype variables available ($n = 167$) for k-means clustering using the flipCluster R package [62].

Statistical analysis

All analyses were completed in R [63]. Code used to generate results are available upon reasonable request to the authors. One BP subject was excluded as an outlier from analyses based on Cook’s distance near 1 in regression models. Correlation analyses and plots were produced using the psych [64] and corrplot [65] packages. Violin plots were produced using the vioplot package [66]. Effect sizes for associations are presented as standardized β coefficients and odds ratios, calculated using the lm.beta [67] R package. The center marks represent median values in the violin plots and boxplots.

Nagelkerke’s R^2 was calculated for variance explained in logistic regression models of disease status by PGS using the RMS R package [68]. All endophenotypes were tested against PGS

and the following covariates in regression models: chromosomal sex, age, self-reported race, self-reported ethnicity, and diagnosis. Protocol/scanner type was controlled for in cognitive control analyses. Overall SZ PGS was included as a covariate for endophenotype analyses by pPGS. Models of continuous variable phenotypes against PGS are represented with partial regression plots.

For cluster versus diagnosis model comparison, we regressed PGSs separately against clusters, plus standard covariates, and against diagnoses, plus covariates. In separate analyses, we regressed PGS against both cluster and diagnosis in the same model, as well as covariates. BP and Cluster 3 were the reference levels for, respectively, diagnosis and cluster.

Multiple testing corrections

Multiple testing corrections were applied for analyses of phenotype, including diagnosis and case status, by pathway PGS. Given the correlated structure of phenotypes and of pPGS in our subjects, we used the poolr R package [69] to calculate the effective number of statistical tests following Galwey (2009) [70]. This methodology identified effectively five (of initially six) independent PGS variables and 12 (of initially 17) independent phenotype variables. Total tested dependent variables were multiplied by total tested independent variables such that findings were corrected for 61 tests (five PGS by 12 endophenotypes plus one test for differential treatment response across k-means clusters) using a Benjamini-Hochberg [71] false discovery rate (FDR) of 0.10.

Permuted null pathway calculation

For each of the pathways in our main analyses, we constructed 10,000 matched null gene sets with an equivalent number of genes, chosen such that the probability of a gene's inclusion is proportional to its length. We then calculated pPGS for these 40,000 permuted null gene sets in

the same way as described above, i.e. using PRSet to calculate pPGS for our target sample and regressing out the first 10 genetic PCs.

Results

Demographics

Subjects were demographically heterogenous and relatively well matched across psychosis cases (n = 205) and controls (*n* = 115). Cases were composed of subjects with SZ, SA, and BP diagnoses (Table 1).

	Cases (*N*)	**(%)**	**Controls (*N*)**	**(%)**
Total	205	64%	115	36%
Diagnosis				
Schizophrenia	118	58%	–	–
Bipolar Disorder	48	23%	–	–
Schizoaffective Disorder	39	19%	–	–
Race				
African American/Black	28	14%	7	6%
American Indian/Alaskan Native	2	1%	0	0%
Asian	15	7%	27	23%
Caucasian/White	131	64%	61	53%
Native Hawaiian/Pacific Islander	3	1%	1	1%
Multiple or unknown	26	13%	19	17%
Ethnicity				
Hispanic/Latino	50	24%	17	15%
Non-Hispanic/Latino	151	74%	97	84%
Unknown	4	2%	1	1%
Inferred 1000 Genomes superpopulation				

African (AFR)	58	28%	19	17%
American (AMR)	20	10%	8	7%
East Asian (EAS)	14	7%	26	23%
European (EUR)	110	54%	53	46%
South Asian (SAS)	3	1%	9	8%
Sex				
Male	158	77%	67	58%
Female	47	23%	48	42%
Mean age (s.d.)	19.7 (4.1)		19.7 (4.2)	

Table 1 Demographics of study subjects.

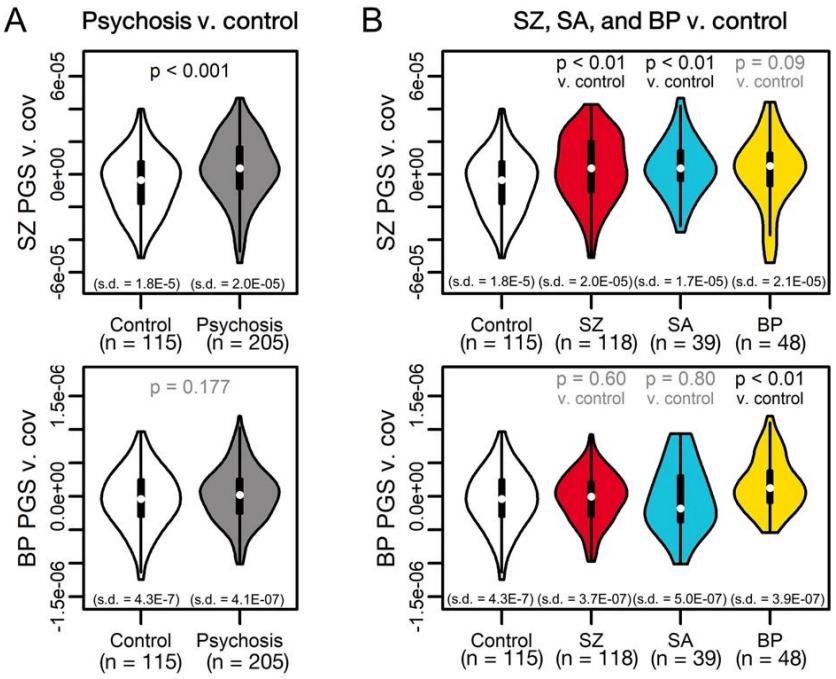
Study subjects were racially and ethnically diverse across both cases and controls. Potential differences in demographics were considered when controlling for covariates in subsequent analyses.

s.d. standard deviation.

Overall and pathway PGSs predict diagnostic status

We first tested whether overall PGS was associated with psychosis status, correcting for latent ancestry differences via inclusion of the top 10 PCs, as well as sex, age, and self-reported race and ethnicity. SZ PGS was associated with case status (OR = 2.69 [CI: 2.06, 3.51]; $p = 1.1 \times 10^{-4}$) (Fig. 1A) and explained 6.3% of variance in status ($p = 1.1 \times 10^{-4}$), similar to previous reports [72,73,74,75,76] (Fig. S1). This association was stronger in SZ (OR = 3.06 [CI: 2.26, 4.14]; $p = 2.1 \times 10^{-4}$) and SA (OR = 3.67 [CI: 2.32, 5.81]; $p = 4.7 \times 10^{-3}$) (Fig. 1B). While BP PGS was not significant in cases overall, it was elevated in BP subjects (OR = 3.55 [CI: 2.34, 5.37]; $p = 2.3 \times 10^{-3}$) (Fig.1A, B). These findings passed multiple testing corrections at FDR < 0.10. SZ and BP PGS were moderately correlated in SZ and SA subjects, but not BP

subjects (Fig. 1C). These results were generally consistent across PGS clumping parameters (see “Methods”, Fig. S2) and when subsetting our cohort to only white European subjects (Fig. S3). Thus, overall PGS captured both transdiagnostic and diagnosis-specific genetic risk for psychotic disorders, consistent with a strong literature on partial genetic overlap between SZ, SA, and BP [6, 7, 77,78,79].



C

Correlations of SZ and BP PGS in cases	All psychosis	SZ only	SA only	BP only
Spearman’s rho	0.30	0.37	0.47	0.00
p-value	< 0.01	< 0.01	< 0.01	0.98

Fig. 1: SZ and BP PGS are associated with psychotic disorder diagnoses.

(A) SZ PGS (top) is significantly associated with psychosis case status, while BP PGS (bottom) is not. (B) SZ PGS (top) is associated with SZ and SA status versus controls, but not BP status. Conversely, BP PGS (bottom) is associated with BP status versus controls, but not SZ or SA status. For each (A) and (B), the top row represents a comparison of SZ PGS across diagnostic groups, while the bottom row represents a comparison of BP PGS

across diagnostic groups. © Spearman's correlations of SZ and BP PGS in psychosis cases show moderate correlations between PGS for SZ and SA subjects, but not for BP subjects.

Association of pPGS with phenotypes

As SZ PGS had stronger overall and transdiagnostic association to psychosis in this cohort, we next tested if this risk could be further subset via pPGS associated with major neurotransmitter pathways. To dissect genetic contributions of biological pathways to clinical and neurobiological phenotypes, we calculated pPGS for loci associated with four neurotransmitter systems relevant to psychosis: glutamate [18,19,20,21,22], GABA [19,20,21,22,23,24], dopamine [25,26,27,28], and serotonin [28, 29], allowing overlap in genes annotated to each pathway (Fig. 2A, Table S1). These pPGS showed moderate intercorrelations and correlations with overall PGS (Fig. 2B).

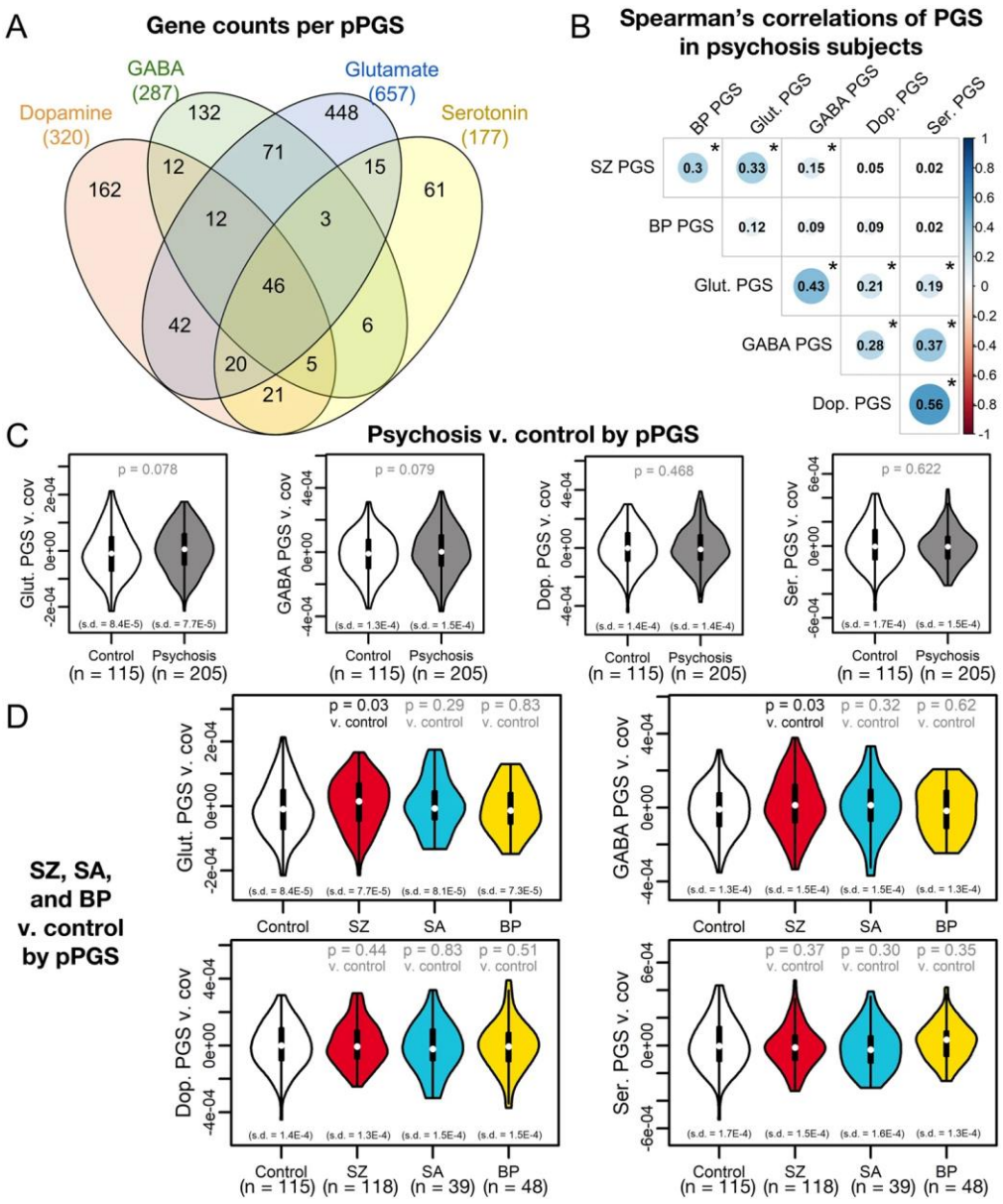


Fig. 2: Pathway polygenic scores (pPGS) based on four main neurotransmitter systems.

(A) KEGG [58], AmiGO [60], and REACTOME [59] were searched for genes associated with glutamatergic, GABAergic, dopaminergic, and serotonergic neurotransmission. (B) pPGS developed from these gene sets showed substantial correlation in psychosis subjects when regressing against the first ten genetic principal components. Statistically significant

correlations ($p < 0.05$) are denoted with an asterisk. (C) While no pPGS significantly separated overall psychosis cases from controls, (D) glutamate and GABA pPGS separated SZ cases specifically from controls, though these findings do not pass FDR < 0.10.

We first tested association with overall disease status. Glutamate and GABA pPGS showed weak trends toward explaining variance in case status (OR = 1.57 [CI: 1.22, 2.04] to 1.58 [CI: 1.22, 2.05]; both $p = 0.08$; Fig. 2C). Glutamate and GABA pPGS showed improved explanation of SZ status over a null background distribution of size-matched gene sets (Fig. S4) and were modestly elevated in SZ subjects relative to controls (OR = 1.90 [CI: 1.42, 2.56] to 1.92 [CI: 1.43, 2.57]; both $p = 0.03$; Fig. 2D), though case-control separation did not pass multiple testing corrections. This suggests disease status in our cohort is partially explained by glutamatergic and GABAergic genetic risk.

pPGSs are associated with endophenotypes in cases

We next tested if PGS was associated with symptom variation across 11 phenotypes within subjects with a psychotic disorder (Fig. 3A). For these analyses, we tested for associations only within psychosis cases, as data on psychosis endophenotypes were only collected for subjects with psychosis. Overall PGSs were not significantly associated with measured phenotypes (Fig. S5). However, we found evidence for associations of pathway-specific risk to six phenotypes in global functioning, role functioning, cognitive control, and treatment response, though only DLPFC β-values by glutamate and GAF by dopamine passed multiple testing corrections at FDR < 0.10. Full data on non-significant relationships is available in Fig. S5.

A

Phenotypes measured

Measure	Phenotype
12-month BPRS change	Treatment response
GAF, GSF, GRF	Global, social, and role functioning
AX-CPT during fMRI	Cognitive control
SANS, SAPS, BPRS	Disorganization
SANS, SAPS, BPRS	Poverty symptoms
SANS, SAPS, BPRS	Reality distortion
YMRS	Mania

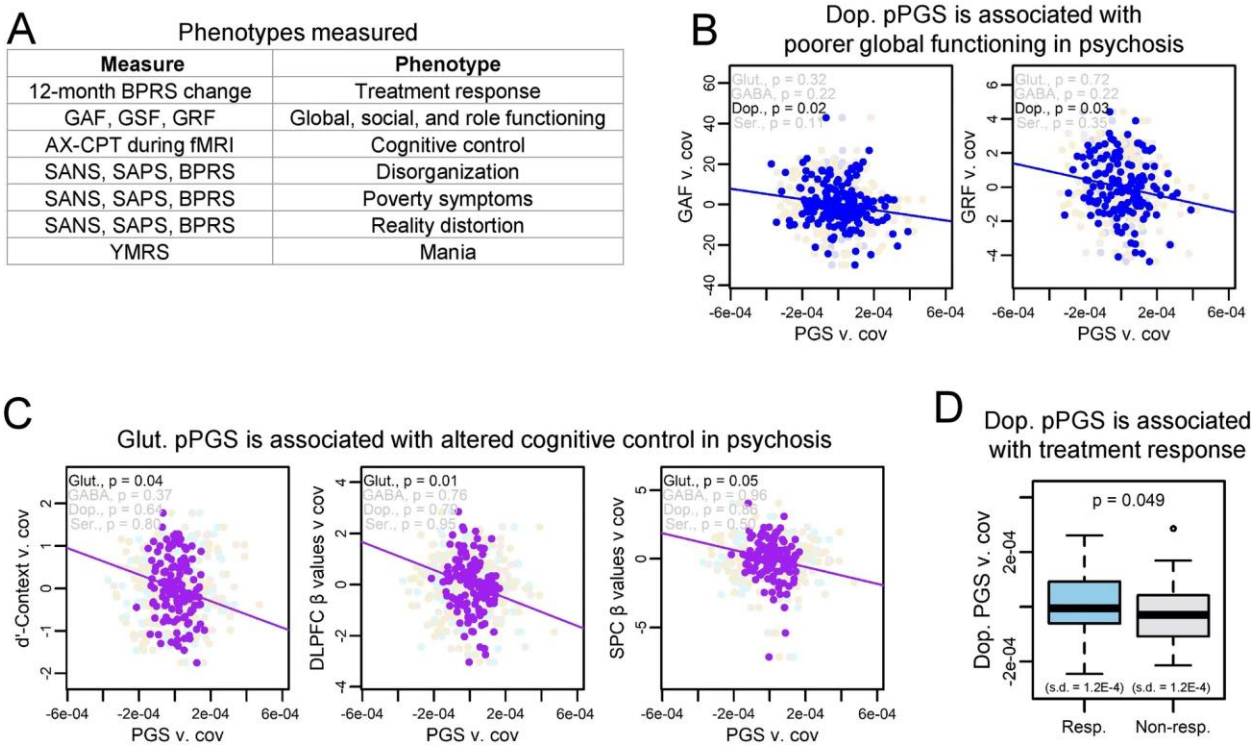


Fig. 3: pPGSs identify mechanisms for endophenotypes of psychotic disorders.

(A) Subjects were phenotyped on a range of psychological, clinical, and neurological measures. (B) Dopamine pPGS was associated with poorer global functioning and role functioning in psychosis subjects when controlling for diagnosis and overall SZ PGS. Only the association with global functioning passed FDR < 0.10. (C) Glutamate pPGS showed associations with poorer performance and reduced cortical activation in psychosis subjects during the AX-CPT cognitive control task when controlling for diagnosis and overall SZ PGS. Only the association with DLPFC β-values passed FDR < 0.10. (D) Treatment response showed a slight association with increased dopamine pPGS, though this did not pass FDR < 0.10. (Resp. = responder to treatment, non-resp. = non-responder to treatment).

Dopamine pPGS was associated with poorer global ($\beta = -0.17$ [CI: −0.25, −0.10]; $p = 0.02$) and role ($\beta = -0.19$ [CI: −0.28, −0.10]; $p = 0.03$) functioning, though role functioning did not pass multiple testing corrections (Fig. 3B). Perhaps the strongest finding was an association between

glutamate pPGS and cognitive control in cases, though only the association to DLPFC *β*-values passed multiple testing corrections (DLPFC β-values: $\beta = -0.22$ [CI: −0.31, −0.13]; $p = 0.01$) (SPC *β*-values: $\beta = -0.19$ [CI: −0.29, −0.09]; $p = 0.05$) (d'-Context: $\beta = -0.18$ [CI: −0.26, −0.09]; $p = 0.04$) (Fig. 3C). This finding is consistent with a prominent hypothesis in schizophrenia literature that perturbations to glutamatergic signaling may be associated with deficits in cognitive control [80,81,82,83]. Notably, our findings represent results from a relatively small cohort and should be followed up in larger cohorts and with more detailed mechanistic studies, and only one endophenotype association to glutamate passed multiple testing corrections. However, these results show the promise of using pPGS to partition genome-wide risk into hypothesized mechanistic pathways toward symptoms of psychiatric disease.

Of 63 subjects with available treatment response data, 35 were responders, consistent with reported efficacy rates [84]. Though underpowered, treatment response showed a weak association with higher dopamine pPGS which did not pass multiple testing corrections (OR = 3.40 [CI: 1.83, 6.34]; $p = 0.05$) (Fig. 3D). While this finding should be considered preliminary, this is consistent with literature associating dopamine with antipsychotic efficacy.

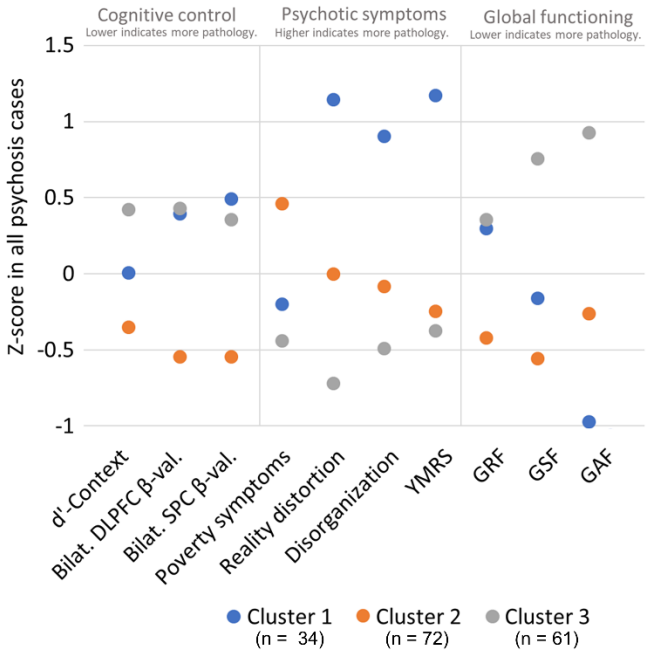
Unbiased phenotype clustering of psychosis cases and genetic burden

Biotype-level grouping of psychosis cases, such as by B-SNIP [9], may advance the use of biological data to better model psychopathology. We hypothesized a biotype-style approach might unmask explanations of psychosis symptom heterogeneity by overall and pathway-level PGS. Toward this goal, we used k-means clustering to group endophenotype presentation among cases. We clustered subjects on z-score normalized phenotype data into three groups following cluster optimization using the NbClust R package [61].

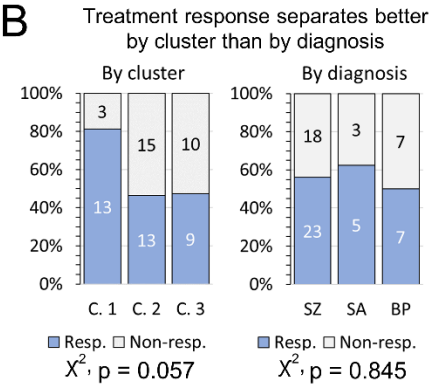
Clusters showed distinct symptom profiles (Fig. 4A, Table S2). Cluster 1 was distinguished by high mania, disorganization, and reality distortion, and moderate impairments in cognitive

control. Cluster 2 had deficits in negative symptoms, showed poor outcomes in role and social functioning, and had the poorest measures of cognitive control. Cluster 3 had low pathology. Cluster 1 showed evidence of enrichment for treatment responders relative to Clusters 2 and 3, though this did not reach significance ($p = 0.057$) (Fig. 4B). In contrast, treatment response showed no separation by diagnosis ($p = 0.845$). Clusters were demographically and diagnostically heterogenous (Fig. 4C, Table S2), though Cluster 3 captured most BP subjects. By contrast, SZ and SA subjects were relatively evenly distributed across Clusters 1 and 2. This suggests that, among endophenotypes available in our subjects, Clusters 1 and 2 captured relatively strong endophenotype differences that are transdiagnostic across SZ and SA subjects.

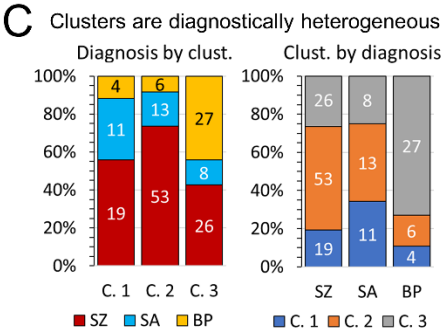
A



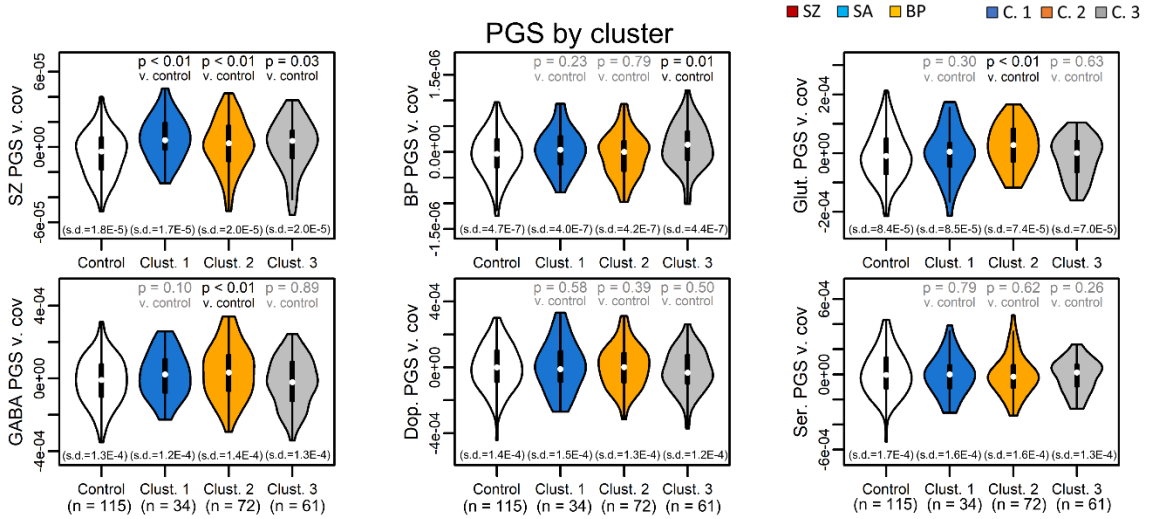
B



C



D



E

Model performance of PGS by cluster v. PGS by diagnosis

	BIC for PGS by:		R^2 for PGS by:	
	Cluster	Diagnosis	Cluster	Diagnosis
SZ PGS	-3086	-3086	0.04	0.04
BP PGS	-4378	**-4380**	0.10	**0.12**
Glut. pPGS	**-2640**	-2634	**0.08**	0.05
GABA pPGS	**-2435**	-2431	**0.05**	0.02
Dop. pPGS	**-2461**	-2460	**0.09**	0.08
Ser. pPGS	-2415	**-2416**	0.05	**0.06**

Better values indicated in **bold**.

F

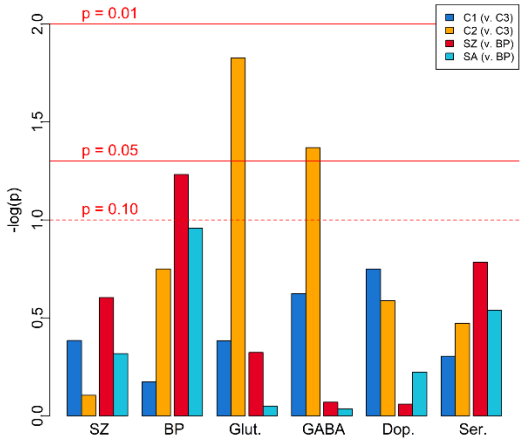


Fig. 4: Psychosis subjects separate into clusters based on symptom profiles.

(A) Psychosis subjects separate into three clusters that are characterized by (Cluster 1) mania, disorganization, reality distortion, and moderate impairments in cognitive control; (Cluster 2) high poverty symptoms, poor social functioning, and strong deficits in cognitive control; and (Cluster 3) overall low levels of pathology. (B) Cluster 1 showed enrichment for treatment responders, though this did not pass FDR < 0.10. In contrast, treatment efficacy did not separate by diagnosis. (C) Diagnoses are mixed across clusters. (D) SZ PGS is elevated in all clusters, though this did not pass FDR < 0.10 for Cluster 3. BP PGS is elevated in Cluster 3. Glutamate and GABA pPGS are elevated in Cluster 2. (E) Models regressing PGS against cluster perform better for glutamate and GABA pPGS than models regressing PGS against diagnosis. Models regressing PGS by diagnosis perform better than PGS by cluster for BP PGS. Weaker evidence for performance differences by model type are present for dopamine pPGS and serotonin pPGS. (F) Glutamate and GABA pPGS are significantly associated with cluster membership when controlling for diagnosis. (C. 1 = Cluster 1, C. 2 = Cluster 2, C. 3 = Cluster 3, resp. = responder to treatment, non-resp. = non-responder to treatment).

We next tested associations between cluster and PGS (Fig. 4D, Table S2). SZ PGS was associated with all clusters, though this did not pass multiple testing corrections for Cluster 3 (Clust. 1: OR = 4.90 [CI: 3.01, 7.96]; $p = 1.1 \times 10^{-3}$) (Clust. 2: OR = 2.59 [CI: 1.83, 3.66]; $p = 6.1 \times 10^{-3}$) (Clust. 3: OR = 2.21 [CI: 1.53, 3.19]; $p = 0.03$). BP PGS was strongly associated with Cluster 3 (OR = 2.51 [CI: 1.74, 3.60]; $p = 0.01$), consistent with its preponderance of BP subjects. Cluster 3 was not associated with any pPGS, suggesting genetic risk may lie outside of SZ variants in neurotransmitter-linked loci. In contrast, GABA (OR = 2.51 [CI: 1.78, 3.53]; $p = 0.01$) and glutamate (OR = 3.21 [CI: 2.24, 4.59]; $p = 1.15 \times 10^{-3}$) pPGS were elevated in Cluster 2, both of which passed multiple testing corrections. This may have implications for the

symptom profiles of this cluster. Cluster 2 showed strong negative symptoms, which are thought to largely arise from glutamatergic dysfunction [85, 86]. Cluster 2 also showed deficits in social and role functioning, which are predicted by negative symptoms [87], and strong cognitive control deficits. pPGS findings suggest Cluster 2's symptoms in our cohort may be more specific to GABAergic and glutamatergic risk, consistent with the hypothesis of an imbalance between these neurotransmitters in psychosis [20]. Following up on these findings in larger cohorts may eventually prove useful in identifying pathways underlying transdiagnostic psychiatric symptoms.

We next compared explanatory power of our clusters versus diagnoses to model genetic risk in our cohort. For this, we defined models regressing PGSs against either cluster or diagnosis. Per Bayesian information criteria and R^2 (Fig. 4E), models using clusters performed better than models using diagnosis to predict, in particular, glutamatergic and GABAergic pPGS. We next regressed PGS against both cluster and diagnosis as covariates to test independence between diagnosis and cluster (Fig. 4F). Again, glutamate ($\beta = 0.23$ [CI: 0.14, 0.33]; $p = 0.02$) and GABA ($\beta = 0.20$ [CI: 0.10, 0.29]; $p = 0.04$) pPGS were associated with Cluster 2 when including diagnosis in the model. These results demonstrate that, within subjects of the same diagnosis, endophenotypes captured here showed specific associations to glutamatergic and GABAergic pPGS. These results provide evidence for the biological validity of these clusters overall within our cohort and within the context of molecular pathways.

Discussion

We tested how overall and pathway-specific PGS are related to diagnosis and variation in psychosis symptoms in a diverse patient sample. We calculated overall SZ and BP PGS, and further curated gene sets representing four major neurotransmitter systems [58,59,60], comparing pPGS for SZ variants in these pathways. Overall genome-wide PGS, and to a

weaker degree glutamate and GABA pPGS, were associated with case status, though glutamate and GABA did not pass multiple testing corrections. pPGSs outperformed overall PGS in explaining endophenotypes. Dopamine was significantly associated with poorer overall functioning and marginally associated with treatment response, though the association to treatment response did not pass multiple testing corrections. This is a preliminary but intriguing result, suggesting the hypothesis for future research that patients with increased dopaminergic risk may be better candidates for antipsychotic medications. Our results also showed associations between glutamate pPGS and cognitive control. Unsupervised clustering identified three phenotypically distinct groups of psychosis subjects that differed predominantly on positive symptoms, negative symptoms, cognitive control, and global functioning. Though our data largely represent different phenotypic measurements than those used by B-SNIP, we reproduced those findings which our measures can capture and generated preliminary hypotheses on the relationship between biotypes and genetic burden. Like B-SNIP, we identified three diagnostically mixed endophenotypes in our subjects that separate on general impairment and cognitive control. Notably, Cluster 2, which showed primary deficits in cognitive control, negative symptoms, and social functioning, had elevated glutamatergic and GABAergic risk. Treatment efficacy showed some evidence for separation by cluster. Within diagnoses, clustering on available endophenotypes unmasked different symptom presentations that showed specific profiles of genetic risk. Our results require replication before generalization outside of our small cohort, but may serve as proof-of-principle for the use of biologically partitioned polygenic risk to study shared mechanisms behind psychiatric symptoms.

Our study features limitations. Notably, our cohort is smaller than many genomics studies, though our deep level of phenotyping is a strength and generally not possible for the larger or combined cohorts needed for GWAS. As a result of our small sample, our results should be considered preliminary data and are best understood as a model for the use of pPGS to study

hypothesized mechanisms of disease. Additional cohorts featuring endophenotypes not captured in our study, such as mood symptoms or potential biomarkers of disease, may help draw a more complete picture of how genetic burden drives psychiatric symptomatology. Focusing on PGS rather than SNP discovery mitigates the limitations of our sample size, though limits our study to integration of previously defined associations. We tested *a priori*-defined and hypothesis-driven pathways in the present study, however, there are certain to be other contributing pathways not investigated here. While diverse cohorts are needed to capture generalized genetic risk across populations, there is currently no standard for transposing PGS across ancestries. Here, we regressed PGS on the first 10 genetic PCs and additionally controlled for race and ancestry in association analyses and show that our methods generally address ancestry differences in PGS in our cohort. However, future studies on larger diverse replication cohorts will be needed. Finally, k-means clustering will identify clusters even when the true nature of the data is continuous or cluster differences are not biologically relevant. The identification of unique PGS profiles for our clusters, as well as evidence for by-cluster differences in treatment response, provide auxiliary evidence supporting separate mechanisms driving distinct clusters of primary symptoms within our subjects. PGSs showed associations to specific clusters even within subjects of the same diagnosis. Additionally, we replicate key findings from previous clustering attempts on psychosis subjects [9].

Our results show the utility of overall and pathway-specific PGS to understand genetic burden and interrogate convergent neurobiological mechanisms underlying disease phenotypes. The neurotransmitter systems tested here have been robustly associated with psychotic disorders [18,19,20,21,22,23,24,25,26,27,28,29]. However, their relationships to specific symptoms have not been elucidated. Mechanisms underlying cognitive and negative symptoms in psychosis are particularly poorly understood, despite being some of the strongest predictors of poor outcomes in psychotic disorders [88,89,90,91]. Our results support a role for glutamate dysfunction in

cognitive, as well as negative and social, symptoms in psychosis. This adds to a wide literature hypothesizing that the biological basis of these symptoms may be partly explained by genetic risk driving glutamatergic abnormalities. Our study and findings represent a model for going beyond overall genetic burden in investigation of the genetic components of psychosis, and, in particular, of maximizing power in smaller but deeply characterized cohorts. While our findings will need to be replicated, they offer a promising look into how genetic burden can be partitioned and can explain specific psychosis diagnoses and endophenotypes. Our work is a step towards a more precise understanding of convergent molecular mechanisms underlying psychotic symptoms, which is critical for realizing the promise of stratified medicine and targeted treatments for those suffering from psychotic disorders.

References

1. Moreno-Küstner B, Martín C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. PLoS ONE. 2018;13:e0195687.
2. van Os J, Hanssen M, Bijl RV, Vollebergh W. Prevalence of psychotic disorder and community level of psychotic symptoms: an urban-rural comparison. Arch Gen Psychiatry. 2001;58:663.
3. Subramaniam M, Abdin E, Vaingankar JA, Sambasivam R, Zhang YJ, Shafie S, et al. Lifetime prevalence and correlates of schizophrenia and other psychotic disorders in Singapore. Front Psychiatry. 2021;12:650674.
4. Sharifi V, Amin-Esmaeili M, Hajebi A, Motevalian A, Radgoodarzi R, Hefazi M, et al. Twelve-month prevalence and correlates of psychiatric disorders in Iran: the Iranian Mental Health Survey, 2011. Arch Iran Med. 2015;18:76–84.
5. Phillips MR, Zhang J, Shi Q, Song Z, Ding Z, Pang S, et al. Prevalence, treatment, and associated disability of mental disorders in four provinces in China during 2001-05: an epidemiological survey. Lancet. 2009;373:2041–53.
6. Lee PH, Anttila V, Won H, Feng YCA, Rosenthal J, Zhu Z, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell. 2019;179:1469–1482.e11.
7. Lee PH, Feng YCA, Smoller JW. Pleiotropy and cross-disorder genetics among psychiatric disorders. Biol Psychiatry. 2021;89:20–31.
8. Light G, Greenwood TA, Swerdlow NR, Calkins ME, Freedman R, Green MF, et al. Comparison of the heritability of schizophrenia and endophenotypes in the COGS-1 family study. Schizophr Bull. 2014;40:1404–11.

9. Clementz BA, Sweeney JA, Hamm JP, Ivleva EI, Ethridge LE, Pearlson GD, et al. Identification of distinct psychosis biotypes using brain-based biomarkers. Am J Psychiatry. 2016;173:373–84.
10. Zhang JP, Robinson D, Yu J, Gallego J, Fleischhacker WW, Kahn RS, et al. Schizophrenia polygenic risk score as a predictor of antipsychotic efficacy in first-episode psychosis. Am J Psychiatry. 2019;176:21–8.
11. Shafee R, Nanda P, Padmanabhan JL, Tandon N, Alliey-Rodriguez N, Kalapurakkel S, et al. Polygenic risk for schizophrenia and measured domains of cognition in individuals with psychosis and controls. Transl Psychiatry. 2018;8:78.
12. He Q, Jantac Mam-Lam-Fook C, Chaignaud J, Danset-Alexandre C, Iftimovici A, Gradels Hauguel J, et al. Influence of polygenic risk scores for schizophrenia and resilience on the cognition of individuals at-risk for psychosis. Transl Psychiatry. 2021;11:518.
13. Choi SW, García-González J, Ruan Y, Wu HM, Porras C, Johnson J, et al. PRSet: pathway-based polygenic risk score analyses and software. PLoS Genet. 2023;19:e1010624.
14. Yao Y, Guo W, Zhang S, Yu H, Yan H, Zhang H, et al. Cell type-specific and cross-population polygenic risk score analyses of MIR137 gene pathway in schizophrenia. iScience. 2021;24:102785.
15. Cosgrove D, Harold D, Mothersill O, Anney R, Hill MJ, Bray NJ, et al. MiR-137-derived polygenic risk: effects on cognitive performance in patients with schizophrenia and controls. Transl Psychiatry. 2017;7:e1012.
16. Cosgrove D, Mothersill DO, Whitton L, Harold D, Kelly S, Holleran L, et al. Effects of MiR-137 genetic risk score on brain volume and cortical measures in patients with schizophrenia and controls. Am J Med Genet B Neuropsychiatr Genet. 2018;177:369–76.
17. Reay WR, Geaghan MP, Atkins JR, Carr VJ, Green MJ, Cairns MJ. Genetics-informed precision treatment formulation in schizophrenia and bipolar disorder. Am J Hum Genet. 2022;109:1620–37.
18. Ohi K, Hashimoto R, Ikeda M, Yamamori H, Yasuda Y, Fujimoto M, et al. Glutamate networks implicate cognitive impairments in schizophrenia: genome-wide association studies of 52 cognitive phenotypes. Schizophr Bull. 2015;41:909–18.
19. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016;19:1442–53.
20. Koshiyama D, Kirihara K, Tada M, Nagai T, Fujioka M, Ichikawa E, et al. Electrophysiological evidence for abnormal glutamate-GABA association following psychosis onset. Transl Psychiatry. 2018;8:211.
21. Reddy-Thootkur M, Kraguljac NV, Lahti AC. The role of glutamate and GABA in cognitive dysfunction in schizophrenia and mood disorders – A systematic review of magnetic resonance spectroscopy studies. Schizophr Res. 2022;249:74–84.
22. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022;604:502–8.

23. Devor A, Andreassen OA, Wang Y, Mäki-Marttunen T, Smeland OB, Fan CC, et al. Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia. Mol Psychiatry. 2017;22:792–801.
24. Karolinska Schizophrenia Project (KaSP) Consortium, Orhan F, Fatouros-Bergman H, Goiny M, Malmqvist A, Piehl F, et al. CSF GABA is reduced in first-episode psychosis and associates to symptom severity. Mol Psychiatry. 2018;23:1244–50.
25. Seeman P, Chau-Wong M, Tedesco J, Wong K. Brain receptors for antipsychotic drugs and dopamine: direct binding assays. Proc Natl Acad Sci USA. 1975;72:4376–80.
26. Creese I, Burt DR, Snyder SH. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. Science. 1976;192:481–3.
27. Seeman P, Schwarz J, Chen JF, Szechtman H, Perreault M, McKnight GS, et al. Psychosis pathways converge via D2High dopamine receptors. Synapse. 2006;60:319–46.
28. Richtand NM, Welge JA, Logue AD, Keck PE, Strakowski SM, McNamara RK. Dopamine and serotonin receptor binding and antipsychotic efficacy. Neuropsychopharmacology. 2007;32:1715–26.
29. Baou M, Boumba VA, Petrikis P, Rallis G, Vougiouklakis T, Mavreas V. A review of genetic alterations in the serotonin pathway and their correlation with psychotic diseases and response to atypical antipsychotics. Schizophr Res. 2016;170:18–29.
30. First M, Spitzer R, Gibbon M, Williams J. Structured clinical interview for DSM-IV-TR axis I disorders, research version, Patient Edition. (SCID-I/P). In: Biometrics research. New York: New York State Psychiatric Institute; 2002.
31. American Psychiatric Association. Global assessment of functioning (GAF). In: Diagnostic and statistical manual on mental disorders, third ed text rev. Washington, DC: American Psychiatric Association; 1987.
32. Auther A, Smith C, Cornblatt B. Global functioning: social scale (GF: Social). Glen Oaks, New York: Zucker-Hillside Hospital; 2006.
33. Niendam T, Bearden C, Johnson J, Cannon T. Global functioning: role scale (GF: Role). Los Angeles: University of California, Los Angeles; 2006.
34. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry. 1978;133:429–35.
35. Andreasen NC. The scales of negative (SANS) and positive (SAPS) symptoms. Iowa City: University of Iowa; 1984.
36. Overall JE, Gorham DR. The brief psychiatric rating scale. Psychol Rep. 1962;10:799–812.
37. Barch DM, Carter CS, MacDonald AW, Braver TS, Cohen JD. Context-processing deficits in schizophrenia: diagnostic specificity, 4-week course, and relationships to clinical symptoms. J Abnorm Psychol. 2003;112:132–43.
38. Howes OD, Bose SK, Turkheimer F, Valli I, Egerton A, Valmaggia LR, et al. Dopamine synthesis capacity before onset of psychosis: a prospective [^{18} F]-DOPA PET imaging study. Am J Psychiatry. 2011;168:1311–7.
39. Howes OD, McCutcheon R, Agid O, de Bartolomeis A, van Beveren NJM, Birnbaum ML, et al. Treatment-resistant schizophrenia: Treatment Response and Resistance in

Psychosis (TRRIP) Working Group consensus guidelines on diagnosis and terminology. Am J Psychiatry. 2017;174:216–29.
40. Smucny J, Lesh TA, Carter CS. Baseline frontoparietal task-related BOLD activity as a predictor of improvement in clinical symptoms at 1-year follow-up in recent-onset psychosis. Am J Psychiatry. 2019;176:839–45.
41. Rosvold HE, Mirsky AF, Sarason I, Bransome ED, Beck LH. A continuous performance test of brain damage. J Consult Psychol. 1956;20:343–50.
42. Henderson D, Poppe AB, Barch DM, Carter CS, Gold JM, Ragland JD, et al. Optimization of a goal maintenance task for use in clinical applications. Schizophr Bull. 2012;38:104–13.
43. Minzenberg MJ, Laird AR, Thelen S, Carter CS, Glahn DC. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. Arch Gen Psychiatry. 2009;66:811.
44. MacDonald AW, Cohen JD, Stenger VA, Carter CS. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science. 2000;288:1835–8.
45. Infinium® Genotyping Data Analysis: a guide for analyzing Infinium genotyping data using the GenomeStudio® Genotyping Module. https://pre.illumina.com.cn/content/dam/illumina-marketing/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf.
46. Patel H, Hyuck Sang L. GenomeStudio Genotyping QC SOP v.1.6. https://khp-informatics.github.io/COPILOT/GenomeStudio_genotyping_SOP.html.
47. Siva N. 1000 Genomes project. Nat Biotechnol. 2008;26:256.
48. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17:1665–74.
49. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36:e126.
50. Wang K, Chen Z, Tadesse MG, Glessner J, Grant SFA, Hakonarson H, et al. Modeling genetic inheritance of copy number variations. Nucleic Acids Res. 2008;36:e138.
51. Rees E, Walters JTR, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. Analysis of copy number variations at 15 schizophrenia-associated loci. Br J Psychiatry. 2014;204:108–14.
52. Grozeva D, Conrad DF, Barnes CP, Hurles M, Owen MJ, O'Donovan MC, et al. Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. Schizophr Res. 2012;135:1–7.
53. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry. 2011;168:302–16.
54. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021;536:17817–29.

55. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1776.
56. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience 2019;8:giz082.
57. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022;109:12–23.
58. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 1999;27:29–34.
59. Joshi-Tope G. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2004;33:D428–32.
60. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics 2009;25:288–9.
61. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an *R* package for determining the relevant number of clusters in a data set. J Stat Softw. 2014. http://www.jstatsoft.org/v61/i06/.
62. DisplayR. flipCluster: Cluster analysis. R package version 1.1.11 2023. https://rdrr.io/github/Displayr/flipCluster/.
63. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.R-project.org/.
64. Revelle W. psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois: Northwestern University; 2022. https://CRAN.R-project.org/package=psych.
65. Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix. 2021. https://github.com/taiyun/corrplot.
66. Adler D, Kelly ST, Elliot T, Adamson J. vioplot: violin plot. 2022. https://github.com/TomKellyGenetics/vioplot.
67. Behrendt S. lm.beta: add standardized regression coefficients to linear-model-objects. 2023. https://CRAN.R-project.org/package=lm.beta.
68. Harrell FE. rms: Regression modeling strategies. 2023. https://CRAN.R-project.org/package=rms.
69. Cinar O, Viechtbauer W. The poolr package for combining independent and dependent p values. J Stat Softw. 2022. https://www.jstatsoft.org/v101/i01/
70. Galwey NW. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet Epidemiol. 2009;33:559–68.
71. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
72. Perkins DO, Olde Loohuis L, Barbee J, Ford J, Jeffries CD, Addington J, et al. Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. Am J Psychiatry. 2020;177:155–63.
73. Calafato MS, Thygesen JH, Ranlund S, Zartaloudi E, Cahn W, Crespo-Facorro B, et al. Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders. Br J Psychiatry. 2018;213:535–41.

74. Smigielski L, Papiol S, Theodoridou A, Heekeren K, Gerstenberg M, Wotruba D, et al. Polygenic risk scores across the extended psychosis spectrum. Transl Psychiatry. 2021;11:600.
75. Vassos E, Di Forti M, Coleman J, Iyegbe C, Prata D, Euesden J, et al. An examination of polygenic score risk prediction in individuals with first-episode psychosis. Biol Psychiatry. 2017;81:470–7.
76. Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
77. Cardno AG, Owen MJ. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. Schizophr Bull. 2014;40:504–15.
78. ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41.
79. Ruderfer DM, Ripke S, McQuillin A, Boocock J, Stahl EA, Pavlides JMW, et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. Cell. 2018;173:1705–15.e16.
80. Morgenroth E, Orlov N, Lythgoe DJ, Stone JM, Barker H, Munro J, et al. Altered relationship between prefrontal glutamate and activation during cognitive control in people with high trait anxiety. Cortex J Devoted Study Nerv Syst Behav. 2019;117:53–63.
81. Falkenberg LE, Westerhausen R, Craven AR, Johnsen E, Kroken RA, L Berg EM, et al. Impact of glutamate levels on neuronal response and cognitive abilities in schizophrenia. NeuroImage Clin. 2014;4:576–84.
82. Falkenberg LE, Westerhausen R, Specht K, Hugdahl K. Resting-state glutamate level in the anterior cingulate predicts blood-oxygen level-dependent response to cognitive control. Proc Natl Acad Sci USA. 2012;109:5069–73.
83. Bustillo JR, Chen H, Gasparovic C, Mullins P, Caprihan A, Qualls C, et al. Glutamate as a marker of cognitive function in schizophrenia: a proton spectroscopic imaging study at 4 Tesla. Biol Psychiatry. 2011;69:19–27.
84. Leucht S, Leucht C, Huhn M, Chaimani A, Mavridis D, Helfer B, et al. Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: systematic review, bayesian meta-analysis, and meta-regression of efficacy predictors. Am J Psychiatry. 2017;174:927–42.
85. Goff DC, Coyle JT. The emerging role of glutamate in the pathophysiology and treatment of schizophrenia. Am J Psychiatry. 2001;158:1367–77.
86. Gruber O, Chadha Santuccione A, Aach H. Magnetic resonance imaging in studying schizophrenia, negative symptoms, and the glutamate system. Front Psychiatry. 2014. http://journal.frontiersin.org/article/10.3389/fpsyt.2014.00032/abstract.
87. Ereshefsky S, Zarubin V, Tryon V, Lesh T, Carter C, Niendam T. T76. Retrospective adhd symptoms in early psychosis: relation to current clinical symptoms and psychosocial functioning. Schizophr Bull. 2020;46:S260–1.

88. Bow-Thomas CC, Velligan DI, Miller AL, Olsen J. Predicting quality of life from symptomatology in schizophrenia at exacerbation and stabilization. Psychiatry Res. 1999;86:131–42.
89. Dickerson F. Social functioning and neurocognitive deficits in outpatients with schizophrenia: a 2-year follow-up. Schizophr Res. 1999;37:13–20.
90. Bowie CR, Reichenberg A, Patterson TL, Heaton RK, Harvey PD. Determinants of real-world functional performance in schizophrenia subjects: correlations with cognition, functional capacity, and symptoms. Am J Psychiatry. 2006;163:418–25.
91. Green MF, Kern RS, Braff DL, Mintz J. Neurocognitive deficits and functional outcome in schizophrenia: are we measuring the “Right Stuff”? Schizophr Bull. 2000;26:119–36.

Acknowledgements

Subject recruitment, blood draws, and phenotyping were supported by NIMH grant R01MH059883. DNA extraction and genotyping were supported by the UC Davis-HKU Collaboration Grant. Analyses and writing were supported by the UC Davis LaMP T32 (NIMH T32MH112507), the NIMH NRSA F31MH129135, and NIGMS grant R35GM119831.

Supplementary figures and tables

Fig. S1: Explanatory power of genome-wide PGS for disease status in our cohort is comparable to that previously reported in the literature.

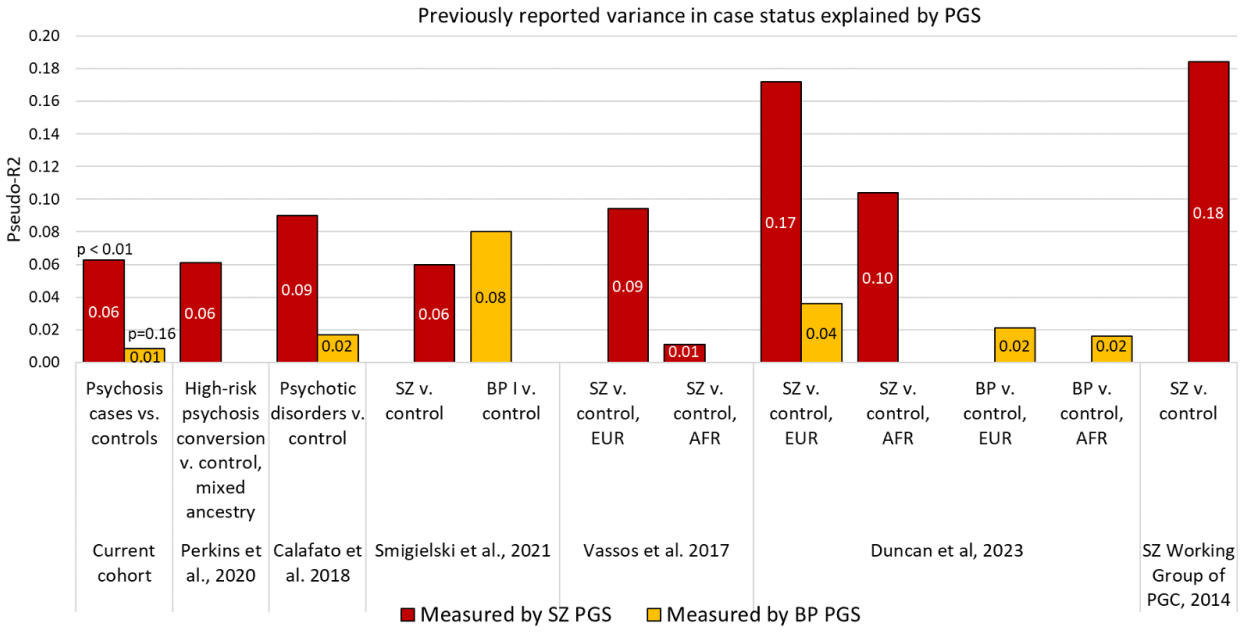
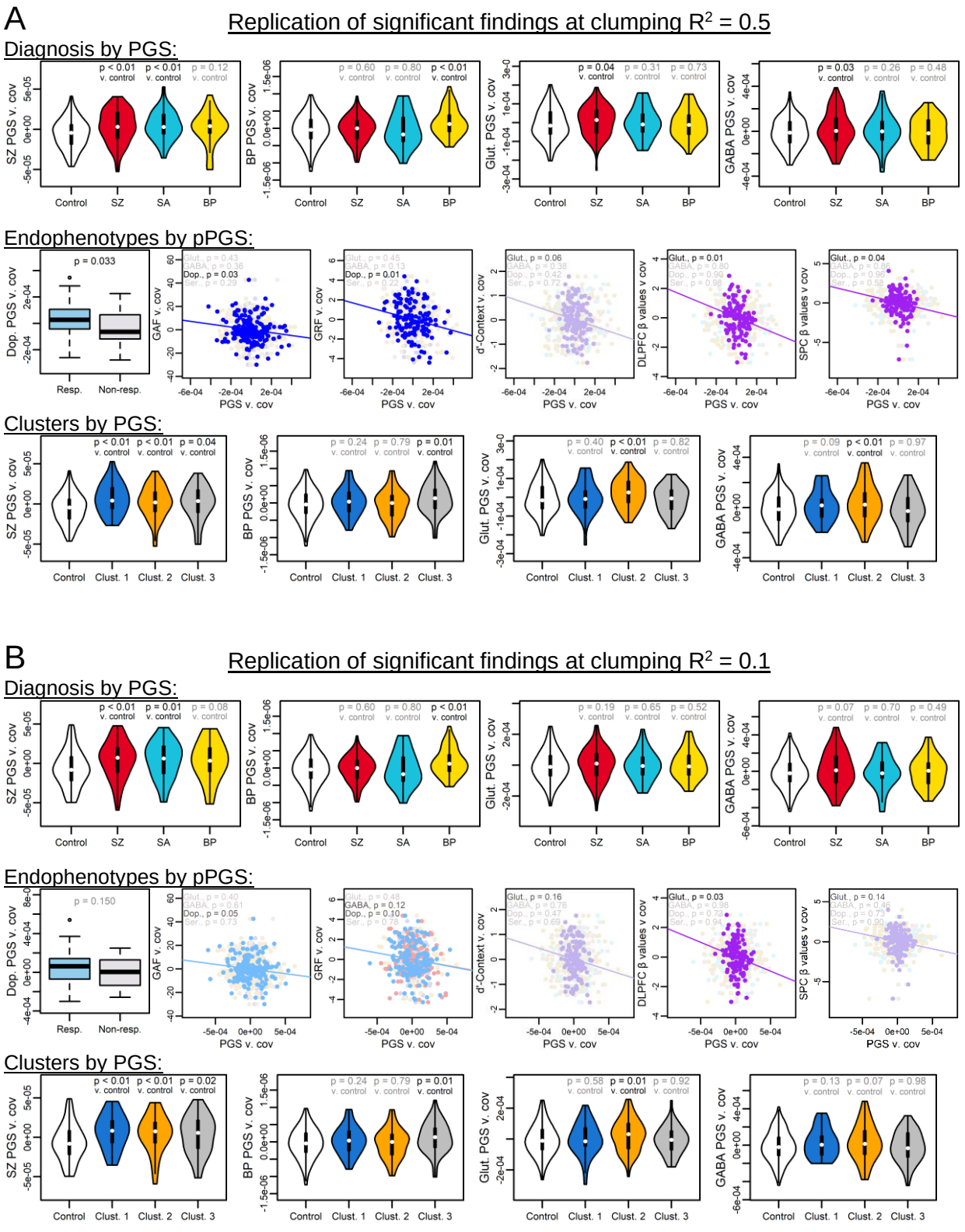


Fig. S2: Significant findings at a clumping LD R^2 of 0.7 show moderate stability at other R^2 values.



C

Replication summary

Finding	p at R^2 = 0.7	p at R^2 = 0.5	p at R^2 = 0.1
SZ vs. control by SZ PGS	< 0.01	**< 0.01**	**< 0.01**
SA vs. control by SZ PGS	0.01	**< 0.01**	**0.01**
BP vs. control by BP PGS	< 0.01	**< 0.01**	**< 0.01**
SZ vs. control by glutamate pPGS	0.03	**0.04**	0.20
SZ vs. control by GABA pPGS	0.03	**0.03**	*0.07*
Treatment response by dopamine pPGS	0.05	**0.03**	0.15
GAF by dopamine pPGS	0.02	**0.01**	*0.05*
GRF by dopamine pPGS	0.03	**0.01**	0.10
d'-Context by glutamate pPGS	0.04	**0.06**	0.16
DLPFC β-value by glutamate pPGS	0.02	**0.01**	**0.03**
SPC β-value by glutamate PGS	0.05	**0.04**	0.14
Clust. 1 vs. control by SZ PGS	< 0.01	**< 0.01**	**< 0.01**
Clust. 2 vs. control by SZ PGS	< 0.01	**< 0.01**	**< 0.01**
Clust. 3 vs. control by SZ PGS	0.03	**0.04**	**0.02**
Clust. 3 vs. control by BP PGS	0.01	**0.01**	**0.01**
Clust. 2 vs. control by glutamate pPGS	< 0.01	**< 0.01**	**0.01**
Clust. 2 vs. control by GABA pPGS	< 0.01	**0.01**	0.07

P-values in **bold text** show strong evidence of replication (less than p ± 0.02 from original result).

P-values in *italic text* show moderate evidence of replication (less than p ± 0.05 from original result).

P-values in blue text show weak evidence of replication (less than p ± 0.10 from original result).

P-values in gray text show little or no evidence of replication (greater than p ± 0.10 from original result).

Fig. S3: Regression against 10 PCs effectively normalizes PGS performance across ancestries in the present cohort. (A) Raw SZ PGS shows strong differences across ancestries, while (B) SZ PGS shows no separation by ancestry within-cases or within-controls after regression. (C) Though underpowered due to additional subsetting of our cohort, main findings show evidence of replication when analyses are constrained to only white/non-Hispanic subjects with estimated EUR ancestry. (Estimated ancestries abbreviated as follows: AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian) (The y-axes of plots A and B are matched for range to improve comparability.)

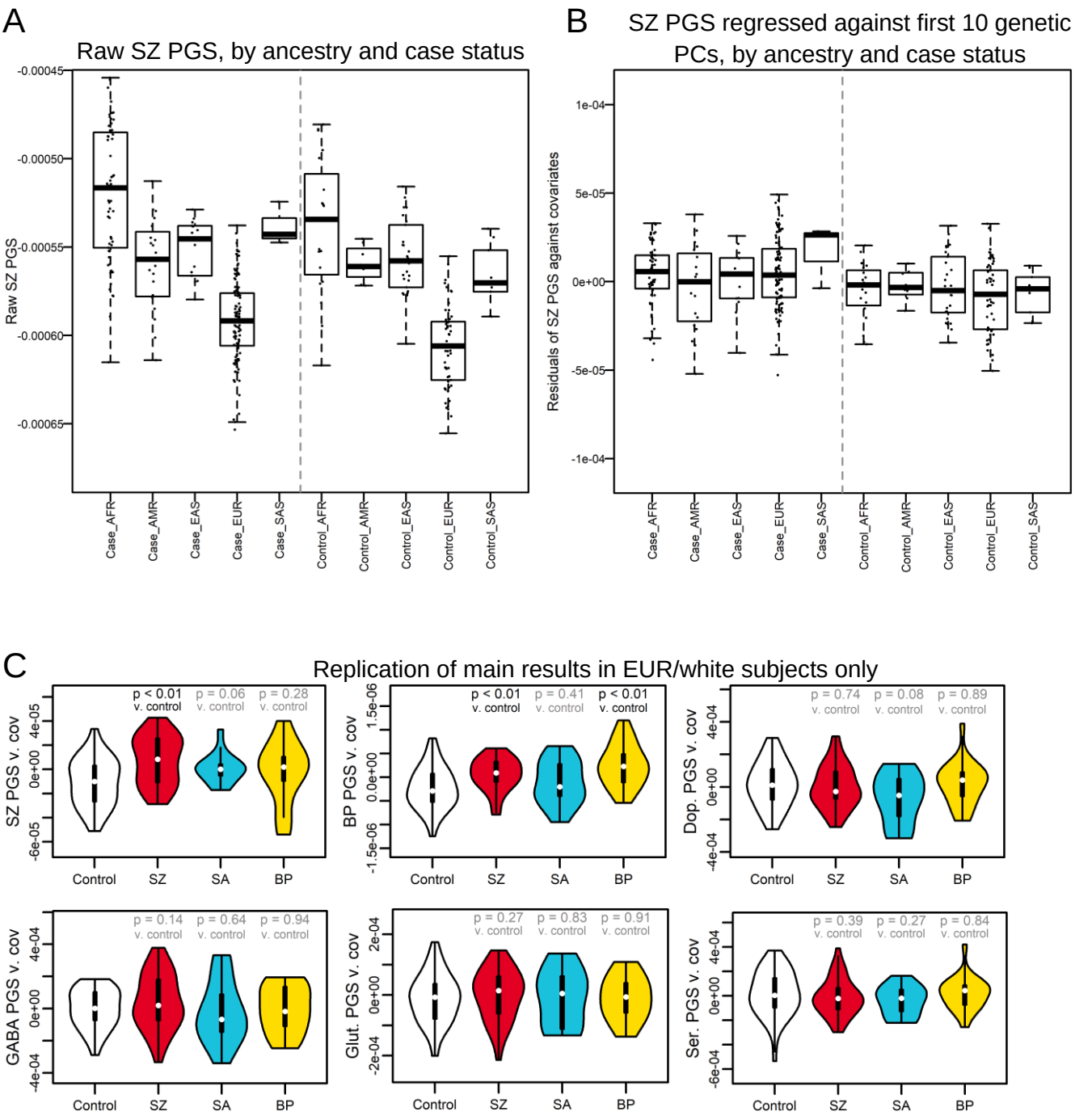

A
Raw SZ PGS, by ancestry and case status
Raw SZ PGS
-0.00045
-0.00050
-0.00055
-0.00060
-0.00065
Case_AFR
Case_AMR
Case_EAS
Case_EUR
Case_SAS
Control_AFR
Control_AMR
Control_EAS
Control_EUR
Control_SAS
B
SZ PGS regressed against first 10 genetic PCs, by ancestry and case status
Residuals of SZ PGS against covariates
1e-04
5e-05
0e+00
-5e-05
-1e-04
C
Replication of main results in EUR/white subjects only
SZ PGS v. cov
p < 0.01 v. control
p = 0.06 v. control
p = 0.28 v. control
BP PGS v. cov
p < 0.01 v. control
p = 0.41 v. control
p < 0.01 v. control
Dop. PGS v. cov
p = 0.74 v. control
p = 0.08 v. control
p = 0.89 v. control
GABA PGS v. cov
p = 0.14 v. control
p = 0.64 v. control
p = 0.94 v. control
Glut. PGS v. cov
p = 0.27 v. control
p = 0.83 v. control
p = 0.91 v. control
Ser. PGS v. cov
p = 0.39 v. control
p = 0.27 v. control
p = 0.84 v. control
Control
SZ
SA
BP

Fig. S4: Glutamate and GABA pPGSs show a weak trend toward partially explaining case status in SZ vs. control subjects, and show evidence of increase over a null background of 10,000 size-matched random gene sets. Dopamine and serotonin PGS show no explanatory power for case vs. control status. Red bar indicates Nagelkerke's R^2 for the neurotransmitter pPGS. pPGSs are plotted against their respective null distributions.

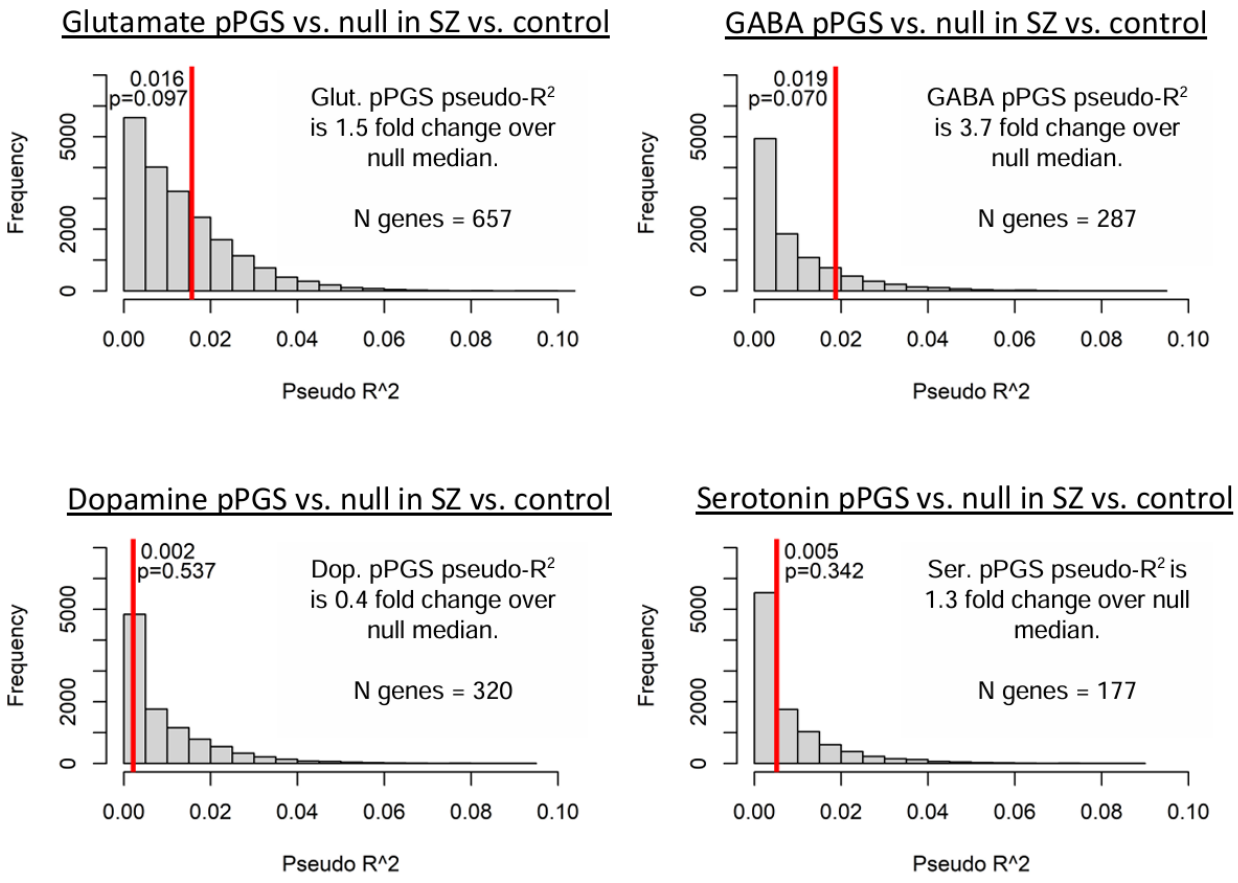


Fig. S5: Phenotypes measured in this study were not significantly associated with overall (A) SZ PGS or (B) BP PGS in our cohort. (C) Non-significant relationships between pPGS and endophenotypes.

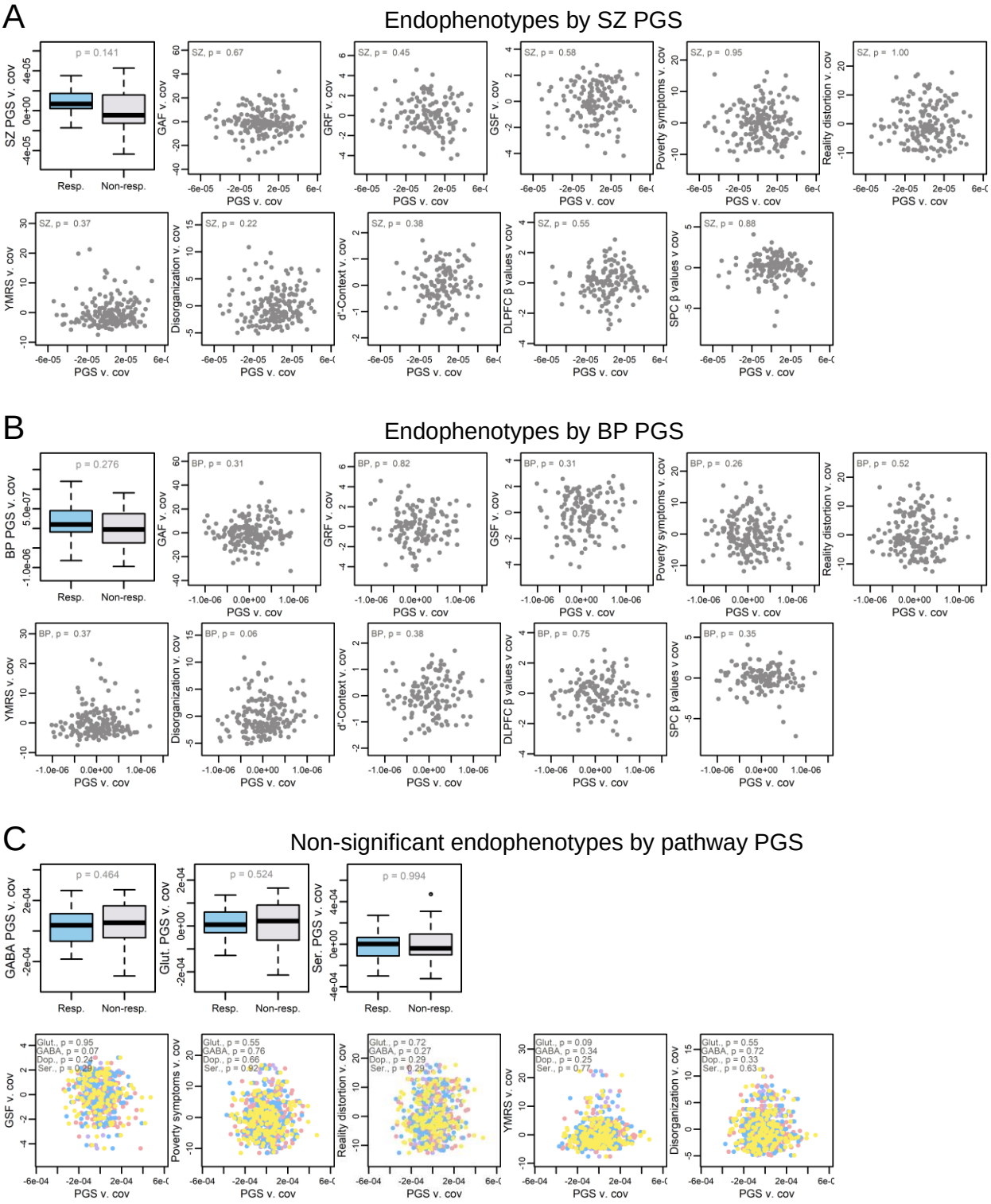


Table S1: Genes in neurotransmitter pathways

Data is available as a downloadable Excel file from https://doi.org/10.1038/s41380-024-02457-0

Table S2: Summary of k-means clusters.

	Cluster 1	Cluster 2	Cluster 3
N	34	72	61
Diagnoses	SZ: 19 (56%) SA: 11 (32%) BP: 4 (12%)	SZ: 53 (74%) SA: 13 (18%) BP: 6 (8%)	SZ: 26 (43%) SA: 8 (13%) BP: 27 (44%)
Primary symptoms	‣ Reality distortion ‣ Disorganization ‣ Mania ‣ Social functioning ‣ Global functioning	‣ Cog. control ‣ Poverty symptoms ‣ Role functioning ‣ Social functioning ‣ Global functioning	(mild deficits)
Highest PGS	‣ SZ PGS	‣ SZ PGS ‣ Glutamate PGS	‣ BP PGS
Race			
African American/Black	3 (9%)	13 (18%)	6 (10%)
American Indian/Alaskan Native	1 (3%)	0 (0%)	1 (2%)
Asian	2 (6%)	8 (11%)	3 (5%)
Caucasian/White	24 (71%)	42 (58%)	41 (67%)
Native Hawaiian/Pacific Islander	0 (0%)	2 (3%)	1 (2%)
Multiple/unknown	4 (12%)	7 (10%)	9 (15%)
Ethnicity			
Hispanic/Latino	10 (29%)	16 (22%)	46 (75%)
Non-Hispanic/Latino	24 (71%)	55 (76%)	14 (23%)
Unknown	0 (0%)	1 (1%)	1 (2%)
Inferred 1000 Genomes SuperPopulation			
AFR	8 (24%)	25 (35%)	12 (20%)
AMR	5 (15%)	7 (10%)	5 (8%)
EAS	1 (3%)	6 (8%)	5 (8%)
EUR	20 (59%)	32 (44%)	39 (64%)
SAS	0 (0%)	2 (3%)	0 (0%)
Sex			
Male	28 (82%)	59 (82%)	45 (74%)
Female	6 (18%)	13 (18%)	16 (26%)
Mean age (s.d.)	20.4 (4.3)	19.1 (4.3)	20.0 (4.0)

Concluding remarks

Psychiatric disorders are currently conceptualized as groups of comorbid symptoms without a precise understanding of their underlying biological mechanisms[1,2]. While this derives from legitimate historical and scientific context as described in my introduction, it has hindered the development of therapies which may target symptoms at their biological roots. To facilitate etiological stratification of patients and therapies, we need an improved understanding of the biological mechanisms which underlie psychiatric symptoms. I aimed to contribute to this work during my PhD by mapping the biological impacts of common psychotic disorder risk variants. I summarize here my research contributions. I subsequently contextualize my work into potential future directions and the wider goal of mapping out convergent etiological pathways.

In Chapters 1 and 2, I used *in vivo* assays to test hypothesized *cis*-regulatory elements (CREs) for activity in postnatal mouse brain. I described detailed methodologies of assay deployment in Chapter 1. I additionally included an appendix featuring experiments where I used these techniques to validate enhancer activity in one of the top schizophrenia risk genes, *CACNA1C*. This work resulted in one first-author publication[3] and my contribution to a main figure and a supplementary figure in a publication on which I am fourth author[4]. I expanded and improved upon these methodologies in Chapter 2. I subsequently deployed a longitudinal MPRA of SZ risk variants in postnatal mouse brain. I identified changes in CRE activity associated with SZ risk SNPs and with adolescent development. Finally, I intersected my results with publicly available datasets to generate hypotheses of targeted genes. These hypotheses should be tested in future research to identify biological pathways which identified CREs may map onto.

In Chapter 3, I described a multi-lab human genomics collaboration which resulted in a co-first author publication[5]. I developed sets of genes which are critical to four major brain synapse types: glutamate, GABA, dopamine, and serotonin synapses. My goal was to test for

relationships between cumulative genetic burden in these pathways and psychotic disorder endophenotypes in humans. My co-first author, Justin Tubbs, calculated pathway polygenic risk scores (pPGSs) for my developed gene sets. I then identified associations in humans with psychotic disorders between neurotransmitter pPGSs and transdiagnostic endophenotypes. Results highlighted associations between glutamatergic and GABAergic dysfunction and deficits in cognitive control and social functioning. I also found associations between dopaminergic dysfunction and improved efficacy of antipsychotic medications. These relationships were not driven by any single disorder, hinting possible transdiagnostic etiologies which may produce shared psychiatric phenomena.

My dissertation research improved our understanding of the biological pathways which link common risk variants to psychiatric symptoms. However, we are still far from a true map of disease etiologies. I believe that future avenues for research can help get us there. For one, my MPRA work only tested for CRE activity. While I generated hypotheses on CREs' targeted genes, these hypotheses remain to be tested. Perturb-seq experiments, which use CRISPR-based tools to disrupt CREs in their native genomic contexts[6], may be a particularly valuable next step for identifying transcriptomic impacts of disruption to CREs. Differentially expressed genes resulting from these experiments could then be analyzed for enrichment of different gene ontologies or drug targets, as was done in Townsley et al (2022)[7]. *In vitro* perturbations could be examined for impacts on cellular phenotypes. These experiments may help us map CRE-disrupting risk variants onto psychiatric etiologies.

Future research using pPGS could test more fine-grained hypotheses than those assessed in my third chapter. For example, researchers could develop pPGSs for specific receptors which may be of interest for drug development. These pPGS could be tested for associations to symptoms or biomarkers. This could even be combined with CRE-testing assays like perturb-seq, using differentially expressed genes as the gene sets for pPGS. Results may illuminate

convergent mechanisms that constitute subgroups of diseases or transdiagnostic pathways.

Information generated from these techniques may be powerful in moving psychiatry toward a stratified mechanism approach. Eventually, it may be possible for clinicians to implement quick, cheap, genetic assays in order to identify patients with high risk in different biological pathways. If so, patients could receive treatments and preventions in a manner which is matched to their individual biologies. My work has not only contributed to the knowledge base that will be required to someday achieve this, but modeled novel approaches that will help facilitate the path toward this goal. My efforts are a part of a wider movement, and I am looking forward to following the research and milestones that will get us there.

References

1. Scadding, J. G. The semantic problems of psychiatry. *Psychol. Med.* **20**, 243–248 (1990).
2. Cooper, J. Disorders are different from diseases. *World Psychiatry* **3**, 24 (2004).
3. Warren, T. L., Lambert, J. T. & Nord, A. S. AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *JoVE* e62650 (2022) doi:10.3791/62650.
4. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10**, e69479 (2021).
5. Warren, T. L. *et al.* Association of neurotransmitter pathway polygenic risk with specific symptom profiles in psychosis. *Molecular Psychiatry* (2024) doi:10.1038/s41380-024-02457-0.
6. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).
7. Townsley, K. G. *et al. Convergent Impact of Schizophrenia Risk Genes*. http://biorxiv.org/lookup/doi/10.1101/2022.03.29.486286 (2022) doi:10.1101/2022.03.29.486286.