Title
Systematic Characterization of Tauopathy-Associated Genetic Risk Loci using Multiplexed Reporter Assays

Author
Cooper, Yonatan

Publication Date
2021

UNIVERSITY OF CALIFORNIA

Los Angeles

Systematic Characterization of Tauopathy-Associated Genetic Risk Loci using Multiplexed Reporter Assays

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Yonatan Cooper

2021

ABSTRACT OF THE DISSERTATION

Systematic Characterization of Tauopathy-Associated Genetic Risk Loci
using Multiplexed Reporter Assays

by

Yonatan Cooper
Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2021
Professor Daniel H Geschwind, Chair

The widespread adoption of genome-wide association studies (GWAS) has revolutionized the detection of genetic loci associated with complex traits. However, the majority of common susceptibility loci reside in poorly annotated noncoding genomic regions and are composed of many correlated polymorphisms due to linkage disequilibrium, obscuring identification of the causal variants and mechanisms underlying trait association. Thus, the functional annotation of noncoding variation is a major impediment to interpretation of genetic risk. Massively Parallel Reporter Assays (MPRA) are a novel experimental approach for the high-throughput functional characterization of noncoding genetic variation, yet remain to be systematically applied to any neurologic disorder. In this dissertation, I utilize MPRA to

characterize variation associated with two neurodegenerative disorders that share tau-protein neuropathology, Alzheimer's disease and Progressive Supranuclear Palsy.

First, I describe the design and implementation of an MPRA to screen 5,706 noncoding variants derived from three GWAS for AD and PSP, identifying 320 regulatory polymorphisms comprising 27 of 34 tested loci. These results enable subsequent identification of novel putative risk genes including *PLEKHM1* and *APOC1* distributed across the complex 17q21.31 and 19q13.32 regions. In Chapter 3, I show that functional predictions from four popular computational algorithms for variant prioritization are discordant both with MPRA results and each other. In Chapter 4, I find that MPRA-defined functional variants preferentially disrupt predicted transcription factor binding sites that converge on enhancers with differential cell-type specific activity in PSP and AD, implicating a neuronal *SP1*-driven regulatory network in PSP pathogenesis. These analyses support a novel mechanism underlying noncoding genetic risk, whereby common genetic variants drive disease risk via their aggregate activity on specific transcriptional programs. In Chapter 5, I perform genome editing to validate four causal loci, identifying *C4* as a novel genetic risk factor for AD. Finally, in Chapter 6, I interrogate technical parameters relevant to assay performance, aiding future studies. Taken together, this work represents a comprehensive characterization of common genetic risk associated with AD and PSP and implicates variants, genes, and transcriptional regulatory networks that represent novel risk factors for neurodegenerative tauopathies.

The dissertation of Yonatan Cooper is approved.

Jason Ernst

Bogdan Pasaniuc

Luis De La Torre-Ubieta

Daniel H Geschwind, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGEMENTS

Chapters 2-6 of this dissertation include materials from the following manuscript currently under review: Cooper, Yonatan; Davis, Jessica E; Kosuri, Sriram; Coppola, Giovanni; Geschwind, Daniel H. “Functional regulatory variants implicate distinct transcriptional networks in dementia”. *In review.* 2021.

Daniel H. Geschwind was the principle investigator, helped conceive of the project, provided funding, and helped write the manuscript. Jessica E. Davis and Sriram Kosuri provided training on the experimental technique and input on experimental design. Giovanni Coppola helped conceive of the project. Yonatan Cooper helped conceive of the project, performed the experiments, performed data analysis, and wrote the manuscript.

Additionally, in Chapter 6, experiments performing nucleofection of MPRA libraries into neural progenitor cells were performed in collaboration with Quiyu Guo PhD in the Geschwind lab. Yonatan Cooper conceived of the experiments and helped with data analysis. Quiyu Guo performed the experiments and collected data.

I wish to extend my deepest thanks to Drs. Daniel H. Geschwind and Giovanni Coppola for their guidance throughout my degree, as well as colleagues and labmates in the Geschwind, Coppola, and Kosuri labs. I would like to thank Sriram Kosuri and Jessica Davis for providing me with excellent technical training in Massively Parallel Reporter Assays, which formed the foundation for this work. I would like to thank my thesis committee composed of Jason Ernst, Bogdan Pasaniuc, and Luis De La Torre-Ubieta for helpful discussion and advice. I would also like to acknowledge the support of the department of Human Genetics at UCLA and the UCLA-Caltech Medical Scientist Training Program.

Finally, I wish to acknowledge my funding sources who provided critical support for this work. I was personally supported by the UCLA-Caltech MSTP training grant T32-GM008042. I was also supported by training fellowship 1F30AG064832 through the National Institute of Aging. This work was further supported by the Rainwater Charitable Foundation award 20180629 as well as grant 5UG3NS104095-04 through the National Institute of Neurological Disorders and Stroke.

VITA

EDUCATION
Bowdoin College, Brunswick, ME
B.A. *Magna cum Laude*, 2013
Majors: Neuroscience, Philosophy. GPA: 3.87 on 4.0 point scale

University of Sydney, Sydney, Australia
Undergraduate Study Abroad Program, Spring 2012

PUBLICATIONS

- **Cooper YA**, Davis JE, Kosuri S, Coppola G, Geschwind DH. Functional regulatory variants implicate distinct transcriptional networks in dementia. *In review.* (2021).

- Sayed FA, Telpoukhovskaia M, Kodama L, Li Y, Zhou Y, Le D, Hauduc A, Ludwig C, Gao F, Clelland C, Zhan L, **Cooper YA**, Davalos D, Akassoglou K, Coppola G, Gan L. Differential effects of partial and complete loss of TREM2 on microglial injury response and tauopathy. *Proceedings of the National Academy of Sciences* **115**, 10172–10177 (2018).

- Lee CYD*, Daggett A*, Gu X, Jiang LL, Langfelder P, Li X, Wang N, Zhao Y, Park CS, **Cooper YA**, Ferando I, Mody I, Coppola G, Xu H, Yang XW. Elevated TREM2 gene dosage reprograms microglia responsivity and ameliorates pathological phenotypes in Alzheimer's disease models. *Neuron* **97**, 1032–1048 (2018).

- **Cooper YA**, Nachun D, Dokuru D, Yang Z, Karydas AM, Serrero G, Yue B; Alzheimer's Disease Neuroimaging Initiative, Boxer AL, Miller BL, Coppola G. Progranulin levels in blood in Alzheimer's disease and mild cognitive impairment. *Annals of clinical and translational neurology* **5**, 616–629 (2018).

- Taniguchi M*, Carreira MB*, **Cooper YA**, Bobadilla AC, Heinsbroek JA, Koike N, Larson EB, Balmuth EA, Hughes BW, Penrod RD, Kumar J, Smith LN, Guzman D, Takahashi JS, Kim TK, Kalivas PW, Self DW, Lin Y, Cowan CW. HDAC5 and its target gene, Npas4, function in the nucleus accumbens to regulate cocaine-conditioned behaviors. *Neuron* **96**, 130–144 (2017).

- **Cooper YA**, Pianka S, Alotaibi NM, Salavati B, Weil AG, Ibrahim GM, Wang AC, Fallah A. Repetitive transcranial magnetic stimulation for the treatment of drug-resistant epilepsy: A systematic review and individual participant data meta-analysis of real-world evidence. *Epilepsia Open* (2017).

- Sørensen AT*, **Cooper YA** *, Baratta MV, Weng FJ, Zhang Y, Ramamoorthi K, Fropf R, LaVerriere E, Xue J, Young A, Schneider C, Gøtzsche CR, Hemberg M, Yin JCP, Maier SF, Lin Y. A robust activity marking system for exploring active neuronal ensembles. *Elife* **5**, e13918 (2016).
  *Co-First Authorship

AWARDED FUNDING

- **1F30AG064832** (**Cooper, PI**) NIH/NIA/UCLA. 9/31/2019 – 6/31/2023
  "Systematic Characterization of Tauopathy-Associated Genetic Variation using Multiplexed Reporter Assays"

INVITED ORAL PRESENTATIONS

- **Cooper, YA.** "Identification of functional regulatory variants implicates distinct transcriptional networks in dementia." UCLA Neurodegenerative diseases Symposium. April 2021. Los Angeles, CA, USA.

- **Cooper, YA**. "Identification of common regulatory variants underlying tauopathies using massively multiplexed assays". *UCLA-Caltech Annual Research Conference*. September 2020. Los Angeles, CA, USA.

- **Cooper, YA**. "Systematic Characterization of Tauopathy-Associated Genetic Variation using Multiplexed Reporter Assays". *UCLA Department of Human Genetics Academic Retreat*. November 2019. Los Angeles, CA, USA.

- **Cooper, YA**. "Systematic Characterization of Tauopathy-Associated Genetic Variation using Multiplexed Reporter Assays". *Turken Research Award and Symposium*. November 2019. Los Angeles, CA, USA.

- **Cooper, YA**. "Characterization of genetic variation using multiplexed reporter assays". *UCLA MSTP Tutorial Series*. February 2019. Los Angeles, CA, USA.

POSTER PRESENTATIONS

- **Cooper YA,** Coppola G, Geschwind D. "Identification of common regulatory variants underlying tauopathies using massively parallel reporter assays**".** *2020 PQG Conference: From Variants and Genes to Clinical Actions*. November 2020. Cambridge, MA, USA.

- **Cooper YA**, Davis J, Kosuri S, Coppola G, Geschwind D. "Systematic Characterization of Tauopathy-Associated Genetic Variation Using Multiplexed Reporter Assays". *UCLA-Caltech Annual Research Conference*. September 2019. Los Angeles, CA, USA.

- **Cooper YA**, Davis J, Jasinska A, Service S, Kosuri S, Freimer N, Coppola G. (October, 2018). "Prioritization of causal variants possibly associated with hippocampal volume using a massively parallel reporter assay." *American Society for Human Genetics Meeting*. October 2018. San Diego, CA, USA.

- **Cooper YA**, Pianka S, Alotaibi NM, Salavati B, Weil AG, Ibrahim GM, Wang AC, Fallah A. "Repetitive transcranial magnetic stimulation for the treatment of drug resistant epilepsy: a systematic review and individual participant data meta-analysis of realworld evidence." *Annual American Epilepsy Society Meeting*. December 2017. Washington D.C. USA.

CHAPTER 1

Introduction:

Functional annotation of noncoding variation and neurodegenerative disease genetics

Can you tell me, Socrates, whether virtue is acquired by teaching or by practice; or if neither by teaching nor practice, then whether it comes to man by nature, or in what other way?

PLATO: MENO

Understanding the heritability of complex traits

Although popularized into its modern formulation by Sir Francis Galton [1], "The Nature vs. Nurture" debate has captured the human imagination for millennia. This maxim summarizes a fundamental question in biology: To what extent do genetic or environmental factors determine observed phenotypic variance? This question is more formally encapsulated by the concept of heritability, defined respectively as the proportion of trait variance in a population arising from either total genetic variation (H^2; broad-sense heritability) or additive genetic variation (h^2; narrow-sense heritability) [2]. Heritability is a population parameter that can be estimated using a variety of schema. Traditionally this has included comparing trait correlations between monozygotic and dizygotic twins [3], or computing regression coefficients of offspring against parental phenotypes, as exemplified by Galton's classic example of hereditary stature [4]. In cases of artificial or natural selection, such as in agricultural settings, h^2 is the regression parameter relating the selection response to the selection differential over multiple generations (breeder's equation; $R = h^{2*}S$). Finally, trait heritability can be computed with either known pedigrees or genetic kinship matrices using linear mixed models to compute variance components while accounting for known fixed effects [5].

Heritability estimation has many downstream applications, informing cross-trait comparisons, artificial selection programs, and power for gene-mapping studies [2]. However, understanding the causal, mechanistic underpinnings of genotype-phenotype relationships requires elucidating *genetic architecture*. This is defined as the number of loci associated with a given trait, as well as the joint distribution of allelic frequencies and effect sizes within these loci [6]. The Genome-Wide Association Study (GWAS) - a modern gene mapping approach - has rapidly proliferated due to the development and plummeting costs of genotyping and sequencing

technologies over the last 15 years. GWAS employ strategically located genetic markers as well as linkage disequilibrium (LD) - the inherited patterns of correlation between genetic variants – to efficiently survey variation across the genome in large population cohorts [7]. While assayed variants are typically common (MAF > 1%), new reference panels, deep imputation, and increased sample sizes have enabled characterization of rarer variation (MAF > 0.1%) [8]. GWAS has proven remarkably successful: as of April 2021 the NHGRI-EBI GWAS repository catalogs more than 254,000 variant-trait associations across more than 4,000 studies [9]. Additionally, Whole Genome Sequencing and Whole Exome Sequencing studies have identified millions of rare and structural polymorphisms across rapidly expanding case-control cohorts [10].

Nevertheless, the biological interpretation of gene-mapping studies remains non-trivial. GWAS survey sets of genomic markers, termed tag-SNPs, chosen for their correlations (i.e. linkage disequilibrium) with larger groups of unmeasured variants. In this way, GWAS does not identify causal variants, as any given association between tag-SNP and trait may actually be indirectly detecting the true causal association of another correlated variant in the LD block. Even when considering sufficiently dense genotyping or imputation, the lead (most significant) SNP may not be the causal variant, due to the presence of multiple causal variants in LD, low power, or other technical factors [11]. Thus, the exploitation of LD in GWAS proves a double-edged sword, aiding discovery power, but hindering interpretability. Underlying causal variants are obscured by the many correlated polymorphisms within loci [7], the majority of which are expected to be functionally neutral [12]. Additionally, the majority of variants from GWAS and NGS studies are identified in noncoding regions and maintain unclear functional relationships to putative target genes. Even more vexing, it is uncertain which cell or tissue types genetic risk might be acting through [7,13] (Challenges summarized in Table 1-1). While there are noteworthy

examples of careful mechanistic validation for individual loci, such as identification of a *C4A* copy number variant underlying the MHC-region association in schizophrenia [14], such work is not feasibly scalable. Thus, at present there is a massive imbalance between identification and biologically interpretation of trait-associated loci and genetic variation. **Therefore, the functional interpretation of genetic variation, particularly noncoding variation, is one of the major challenges facing modern genetics.** Below, I will outline current strategies for annotation and prioritization of noncoding variation, emphasizing methodological advantages and drawbacks. I will then introduce Massively Parallel Reporter Assays, a novel approach for functional characterization of noncoding variation.

Challenge	Exacerbating Variables	Methodological Solutions
Is the variant causal?	Linkage disequilibrium Low power/ small effect sizes Genotyping / imputation failure	Fine-mapping Deep imputation with improved population reference panels, increased sample sizes
Is the variant functional?	Same as above Uncertain function of noncoding genome	Prioritization algorithms Functional genomic annotations Genome editing/ *in vitro* modeling MPRA
Which gene is regulated?	Gene dense regions Strong LD Complex 3D genomic architecture	QTL colocalization 3D interactome assays Genome editing/ *in vitro* modeling
Which cell-type mediates risk?	Trait in heterogenous tissue, or unclear causal organ system Gene dense regions	Heritability enrichment by cell-type Transcriptomics/Proteomics Genome editing/ *in vitro* modeling

Table 1-1. The major challenges in interpreting noncoding loci and variants identified in GWAS and NGS studies. Proposed methodological solutions to address these challenges are also displayed.

Interpretation of noncoding genetic variation using functional genomic maps

Noncoding genetic variation is assumed to primarily function by directly or indirectly influencing gene expression (Figure 1-1). Therefore, an efficient and simple prioritization strategy is to overlap variants with functional genomic annotations. These annotations are identified through empirical assays leveraging next generation sequencing to query DNA regions enriched for specific biochemical modifications (called "peaks") known to correlate with regulatory activity. This includes methods to identify DNA accessibility (DNase-hypersensitivity [15] and ATAC-seq [16]), DNA methylation, histone modifications (Histone ChIP-seq) [17], Transcription Factor Binding (TF-ChIP, HT-SELEX, PBM [18]), or direct assessments of transcriptional activity (GRO-seq [19], PRO-seq [20], CAGE [21]). Large-scale consortium initiatives including NIH Roadmap [22], ENCODE [23], IHEC [24], and FANTOM5 [25] catalog multiple such marks across many cell-lines and tissues, and are an invaluable public resource. Moreover, the development of computational tools including ChromHMM [26] and Segway [27] that integrate multiple annotations across tissues provides higher resolution genomic segmentation and more refined descriptions of transcriptional regulatory states. Other tools, including RegulomeDB [28] and HaploReg [29], integrate pre-existing or user defined marks to directly prioritize GWAS variants. Similarly, functional maps are leveraged by a number of computational algorithms and machine learning methods for functional variant prediction (discussed in Chapter 3). Conceptually, functional variation could act within regulatory regions to modify binding of transcriptional complexes (TFBS disruption), by modifying the chromatin architecture directly

(ex. hQTL, meQTL), or some combination of both (Figure 1-1).

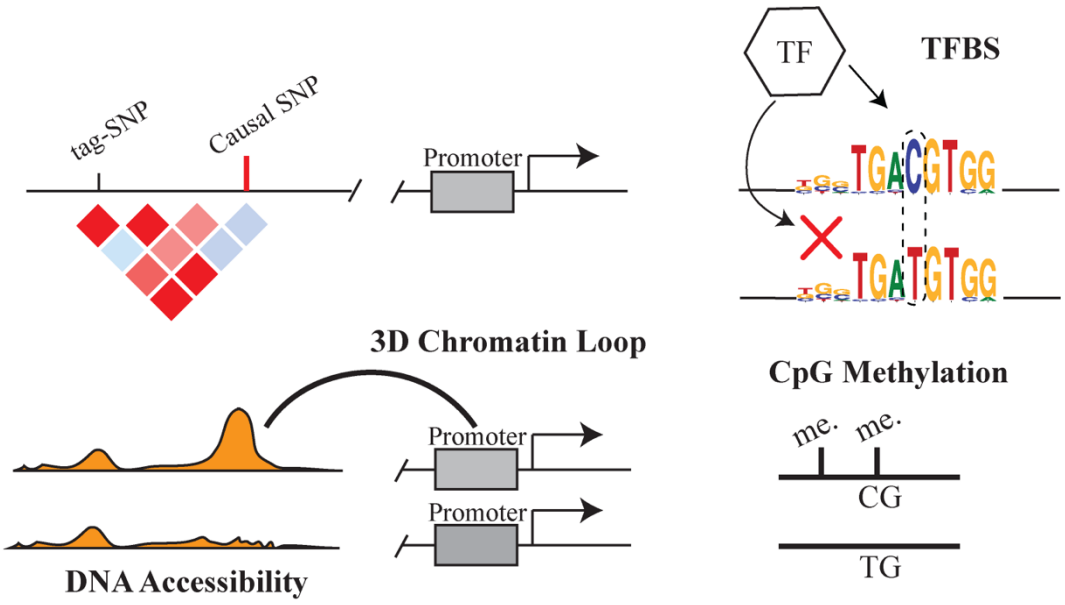


Figure 1-1. Functional mechanisms of transcriptional regulatory variants. Shown is an example trait-associated locus identified by GWAS. The underlying causal variant is in tight linkage with the tag-SNP and is in a noncoding regulatory genomic region. The alternate (T) allele might function by disrupting a binding site for a key transcription factor and directly impacting gene expression. The variant might also change chromatin structure to indirectly influence expression, including through histone modifications and DNA accessibility (hQTL), 3D chromatin looping, or CpG methylation status (meQTL).

The mechanistic relevance of functional genomic annotations to gene regulation and variant prioritization is supported by the strong statistical enrichment of GWAS variation within these regions [30,31]. This has been observed repeatedly, using both simple enrichment tests (such as binomial tests [32]) and more sophisticated regression-based approaches such as stratified LD-Score Regression [33–35], fgwas [36], or GARFIELD [37]. In addition, partitioning of polygenic risk can be used to characterize tissue or cell-type specific enrichment. For example, polygenic risk influencing platelet volume and count was enriched within open chromatin for CD34+ precursor cells, as expected [36]. Such work begins to illuminate the cell-types mediating genetic risk and is especially relevant for traits affecting multiple tissues or heterogenous organs like the brain [38].

Nevertheless, functional genomic maps are subject to a number of limitations. First, peaks are relatively broad and lack the nucleotide-level resolution to disambiguate closely spaced variants [39]. Second, existing functional maps are a static snapshot and may not recapitulate critical demographic, disease, or environmental -specific states, nor dynamic stimulus- induced regulatory changes [40,41]. Moreover, such maps are often derived from immortalized cell-lines or bulk tissues, and are difficult to obtain from rarer cell-types, though single-cell epigenomic studies [42] as well as cross-tissue imputation strategies [43] attempt to rectify this. Third, intergenic regulatory elements and the corresponding GWAS variation they contain require assignment to relevant target genes. This can be addressed through integration with QTL and expression data (discussed below), as well as functional maps of 3D genetic architecture. Modern iterations of chromatin conformation capture, including CHIA-PET, Hi-C, and Hi-ChIP enable genome-scale characterization of the chromatin interactome, physically linking regulatory regions with cognate genes [44,45]. A classic example is the obesity-associated variant rs9930506 located within an intron of *FTO*, which was found to interact and regulate *IRX3* located 1.2 Mb away [46]. Such studies generally highlight that regulatory regions such as enhancers interact with the nearest genes approximately 40% of the time, and often have complex one-to-many and many-to-one interaction relationships [47,48]. However, these techniques are limited by kilobase resolution, moderate sensitivity, and a current lack of comprehensive cell-type and state-specific studies. Finally, and most critically, overlap of GWAS variation with regulatory annotations does not prove causal relationships. Overlap does not necessarily imply functional disruption of regulatory elements or dysregulation of gene expression.

Colocalization with Quantitative Trait Loci for variant prioritization

Quantitative Trait Loci (QTLs) map associations between genotypic variation and quantitative phenotypes, most commonly gene expression (eQTLs) due to the low cost and robustness of transcriptomic measurement. As causal noncoding GWAS variation is expected to regulate gene expression, it is intuitive to interpret GWAS loci through colocalization with eQTLs [49] . Indeed, GWAS variants are significantly enriched for eQTLs compared with background variation [50]. A simple method for integrating GWAS and eQTL data is to simply assess overlap between the two. However this leads to a large proportion of false positives due to LD, considerable genomic pleiotropy, and because a large proportion of variants are eQTLs (estimated at 48%) [51,52]. This motivated the development of statistical methods including COLOC [53], enloc [54], and eCAVIAR [55] that formally test whether two overlapping association signals share common causal variants. These approaches were initially used for fine-mapping and gene assignment for lipid, insulin, and glucose related traits. Large scale eQTL resources, including the Genotype Tissue Expression project [56,57] that curates 54 tissues from 938 individual donors, have greatly facilitated GWAS interpretation. The PsycheENCODE [58], AMP-AD, and Common Mind Consortium provide brain specific resources [59].

An alternative approach to QTL colocalization is the transcriptome wide association study (TWAS), which directly correlates changes in gene expression with downstream traits [60]. Because it is prohibitive to directly measure gene expression from tens of thousands of individuals, a key insight was to leverage eQTL panels to impute gene expression onto GWAS datasets, either for individual-level data [61] or using summary statistics [62]. By collapsing association testing down to thousands of genes rather than millions of SNPS, TWAS reduces the

multiple-testing burden to increase statistical power. Power is also increased in cases of allelic heterogeneity by integrating multiple causal signals into a single expression effect [60].

QTL integration and colocalization provides key advantages in interpretation of GWAS loci. Unlike overlap with genomic annotations, QTL analysis provides the relevant target genes (or relevant quantitative trait). Moreover, integrated colocalization and fine-mapping approaches provide statistical measures of shared causal mechanisms at loci. However, QTLs at present suffer from a number of limitations, most critically in regards to the reference panels used. Even the largest databases, such as GTEx, have only a few hundred individuals per condition, resulting in suboptimal discovery power. In line with this observation, the PsychENCODE consortium required ~1400 samples to nearly saturate discovery of protein-coding eGenes in bulk brain tissue [58]. With the exception of blood, most QTL databases use bulk tissues, which are dominated by the most abundant cell-type, which is problematic in heterogenous tissues [63]. Similarly, there are a dearth of trait or stimulus specific QTLs, and such data would be difficult to generate.

Massively Parallel Reporter Assays enable direct functional characterization of diverse genomic features

The vast majority of GWAS variation and partitioned trait heritability is contained within the noncoding genome, the biological interpretation of which has undoubtedly benefited from the large-scale generation of functional genomic maps across multiple tissues and cell-types [64]. However, such maps are often discordant and fail to overlap functional regulatory elements or variants. For example, a recent analysis in K562 cells found little genomic overlap between multiple different enhancer annotation methods, and found that only a small percentage of

GWAS and eQTL variation (including curated causal variants) overlapped any given enhancer mark [65]. Additionally, experimental characterizations of predicted enhancers find that between 30-46% are transcriptionally active [66,67], which taken together suggests that a large proportion of enhancers are both miss-specified and undetected. Similarly, at present QTL studies remain underpowered, in-line with the observation that only 26% of GWAS loci are explainable by colocalization with eQTLs [68]. These limitations underscore the need for direct and high-throughput functional characterization of noncoding genetic elements and variation within relevant biological contexts, and motivated the development of a diverse set of experimental approaches known as **Massively Parallel Reporter Assays (MPRA)**. MPRA involves the construction of a synthetic library containing a large collection of genomic elements each paired with a reporter gene and a unique, genetically encoded barcode. These libraries are delivered to cell-lines or tissues of interest, and the functional effects of library elements are assayed through the multiplexed measurement of barcoded reporter transcripts using next generation sequencing [69]. This method has enabled regulatory characterization of diverse sets of genomic features across a variety of biological contexts.

The earliest MPRA iterations were developed to test the transcriptional activity of enhancers [68,70–72] . In these assays, putative enhancer elements drive expression of a barcoded reporter gene, and transcriptional activity is assessed as the normalized count of uniquely barcoded transcripts deriving from each element. There are a variety of assay designs, for example: enhancer elements placed upstream of a minimal promoter vs. in the 3' UTR (STARR-seq [73], suRE [74,75]), enhancer DNA obtained via microarray synthesis, PCR [76], or genomic DNA capture [77,78], and episomal vs. integrating assays [79,80] . These assays have been used to successfully screen for enhancer activity [69], repressive elements [81], or differential activity across

a diverse array of prokaryotic and eukaryotic cell types [70–72,82] . Packaging libraries within viral delivery platforms, including Adeno-Associated Viruses (AAV) [78] and Lentivirus [79,83] have broadened the available cellular contexts to include difficult primary cell lines such as neural cells [83,84], and *in vivo* assays in the mouse retina [85] and brain [78]. Approaches incorporating saturation mutagenesis [86] or tiling of overlapping elements [87] enable elucidation of TF-binding logic and nucleotide-resolution sequence specificity. Finally, more recent MPRA iterations probe the architecture of other noncoding genomic features such as 5'and 3' UTRs or splice sites and their effects on transcriptional, post-transcriptional, or translational regulation. These include MPRA characterization of splicing [88,89], RNA or protein stability [90–92], RNA-editing [93], and translation efficiency [94].

MPRA can also be used to compare transcriptional regulatory effects between alleles, though these assays are performed less frequently due to the challenge of measuring small allelic effect sizes. In a landmark study, Tewhey and colleagues characterized more than 32,000 variants associated with LCL eQTLS in K562 cells, identifying ~3400 active regulatory elements and 842 expression modulating variants with significant transcriptional skew between alleles. Of note, 53 of these were well annotated trait-associated variants [95]. Another study assessed 2,756 variants derived from 75 GWAS loci for red blood cell traits, identifying 32 functional expression modifying variants, 3 of which were validated using isogenic genome editing [96]. Additionally, MPRA has been used to measure allelic effects of variants associated with cancer [97,98], osteoarthritis [99] , COPD [100], Lupus [101], and neuropsychiatric disorders [102,103] (Table 1-2). However, it has not been used to test variation associated with any neurologic disorders.

Study	PMID	Trait	# Variants	Cell Type
[96] Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits	27259154	Red Blood Cells	2,756	K562

[95] Direct Identification Of Hundreds Of Expression-Modulating Variants Using A Multiplexed Reporter Assay	27259153	Lymphoblastoid eQTLs	32,373	K562 HepG2
[99] Functional testing of thousands of osteoarthritis-associated variants for regulatory activity	31164647	Osteoarthritis	1,605	Saos-2
[97] Systematic identification of regulatory variants associated with cancer risk	29061142	Cancer	10,763	HEK293T
[98] Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma	32483191	Melanoma	832	UACC903 HEK293FT
[103] Common Genetic Risk Variants Identified In The Spark Cohort Support Ddhd2 As A Candidate Risk Gene For Autism	32747698	Autism Locus	98	HEK293T
[102] A Screen Of 1,049 Schizophrenia And 30 Alzheimer's-Associated Variants For Regulatory Potential	31503409	Schizophrenia	1,049	K562 SK-SY5Y
[101] Global Discovery Of Lupus Genetic Risk Variant Allelic Enhancer Activity	33712590	Lupus	3,093	GM12878
[100] Identification Of Functional Variants In The FAM13A Chronic Obstructive Pulmonary Disease Genome-Wide Association Study Locus By Massively Parallel Reporter Assays	30079747	COPD	606	Beas-2B
[75] High-throughput identification of human SNPs affecting regulatory element activity	31253979	Human variants	5.9 million	K562 HepG2

Table 1-2. Studies utilizing MPRA to screen trait-associated human genetic variation.

Neurodegenerative disease genetics: A brief overview

Neurodegenerative diseases are a clinically and pathologically heterogenous group of disorders characterized by the progressive destruction of nervous system tissues and resultant cognitive, behavioral and motor deficits. Genetics has long been understood to play a fundamental role in disease etiology. Initial studies examining genetic risk found success employing linkage analysis in familial pedigrees of monogenic disorders, exemplified by identification of triplet repeat expansions underlying Huntington's disease [104] and Spinal Cerebellar Ataxia type 1 [105]. This was followed by identification of multiple mendelian risk genes converging on single disease entities, most notably association of mutations in APP and PSEN1/2 with early onset Alzheimer's disease (AD) [106], and SCNA and Parkin with Parkinson's

disease (PD) [107]. Early identification of mendelian risk genes established putative causal mechanisms in turn motivating drug development pipelines. In AD, high-penetrance mutations in proteins responsible for amyloid processing resulted in the "amyloid hypothesis" and a variety of amyloid reducing therapeutics such as aducanumab [108], which looks to receive FDA approval in 2021.

Nevertheless, the underlying genetic architecture and supervening pathophysiology remains complex for the majority of neurodegenerative disorders. For instance, the *C9ORF72* hexanucleotide repeat expansion mysteriously causes FTD in some individuals and ALS in others, disorders with shared pathologies yet dissimilar clinical and anatomical manifestations [109]. Similarly, subtle changes in gene dosage can cause profound phenotypic divergence. For example, single copy mutations in TREM2 confers significant risk for AD, while homozygous loss-of-function results in Nasu-Hakola disease, a rare presenile dementia with bone cysts [110]. Similarly, loss-of-function mutations in PGRN result in familial Frontotemporal dementia (FTD) [111], while small reductions in PGRN expression from the common rs5848 polymorphism increase risk for AD [112]. Lastly, many disorders have complex disease architectures, with completely penetrant, rare high-impact, and common variants, all contributing to risk.

For many neurodegenerative diseases, including AD, PD, FTD-ALS, and Progressive Supranuclear Palsy (PSP), the majority of cases are sporadic, with genetic risk mostly conferred by common polymorphisms. Likewise, polygenic risk has a substantial impact on disease trajectory even in monogenic disorders such as HD [113]. Neurodegenerative disease has a large heritable component, estimated at 60-80% in AD for example [114], and our genetic understanding has greatly benefited from the GWAS revolution [115–121] (Table 1-3). As expected, GWAS of AD and PD (common diseases) benefit from the non-linear relationship between increased sample

size and locus-discovery [122].The latest meta-analyses have identified 69 disease-associated loci from 111,326 AD (or proxy) cases [116] and 90 loci from 56,306 PD cases[117], vs. only 2 loci from a study with 3,526 FTD cases [120]. The proliferation of GWAS across neurodegenerative diseases has provided key mechanistic insights. First, distinct clinicopathologic disorders share substantial heritable risk, pointing at shared underlying disease mechanisms [123]. Moreover, there seems to be shared polygenic risk on the basis of overlapping disease pathology [124], highlighting the imprecision of current clinical diagnostics and motivating an expansion of disease categorization to include genetic and pathologic features [123]. Second, GWAS risk variants and loci enrich within biological pathways and cell types. For example, AD polygenic risk implicates metabolic function, immune signaling, and microglia, findings that have been subsequently validated by functional studies [125,126] .

Nevertheless, for many of the reasons described above, the granular functional interpretation of most neurodegeneration GWAS loci has remained limited. This is highlighted by the MAPT locus, which confers risk for tauopathies including PSP and CBD as well as the α-synucleinopathy PD [127]. How a single locus might confer risk for multiple distinct clinicopathologies might be explained by differing underlying causal variants, haplotypes, or the influence of additional risk genes mediated by LD, but this remains to be explored (Chapter 2).

Study	PMID	Disorder	Cases/Controls*	Loci Identified
Bellenguez et al. 2020	NA	Alzheimer's disease	111,326/401,577	69 (31 new)
Nalls et al. 2019	31701892	Parkinson's disease	56,306/1,417,791	90 (38 new)
Nicolas et al., 2018	29566793	Amyotrophic lateral sclerosis	20,806/59,804	6 (1 new)
Chen et al., 2018	30089514	Progressive Supranuclear Palsy	1,646/10,662	5 (4 suggestive)
Kouri et al., *2015*	26077951	Corticobasal degeneration	219/3,750	2
Ferrari et al., *2014*	24943344	Frontotemporal dementia	3,526/9,402	2
Chia et al., *2021*[&]	33589841	Lewy body dementia	2,591/4027	5

* Cases for AD and PSP GWAS include proxies

[&] Genotyping performed using WGS

Table 1-3. Summarizes largest GWAS to date (# cases) for select neurodegenerative disorders.

Project goals

In summary, neurodegenerative disorders remain one of the most intractable clinical problems today. More than four decades of genetic analyses have uncovered a substantial genetic contribution to neurodegeneration and provided key insights into causal disease biology. Due to recent technical advances, the research community has identified many new degeneration-associated loci and variants. However, the functional interpretation of this variation has remained unclear. Here I will describe the development and application of a massively parallel reporter assay to systematically characterize noncoding variation derived from GWAS for two distinct neurodegenerative disorders, Alzheimer's disease and Progressive Supranuclear Palsy. I will pay particular attention to two complex haplotypes, 17q21.31 and 19q13.32 that are of particular interest to the field, and substantially benefit from empirical functional analysis (Chapter 2). Additionally, I will compare MPRA experimental data with existing computational algorithms (Chapter 3), and discuss how these data can inform empirical assessments of transcription factor binding dysregulation which may play a role in disease pathogenesis (Chapter 4). I will perform gold standard validation on a subset of my predictions using genome editing (Chapter 5). Finally, I will discuss technical factors and experimental consideration that will aid the design and execution of future high-throughput functional approaches (Chapter 6).

Bibliography

1. Galton, F. The history of twins, as a criterion of the relative powers of nature and nurture. *Frasers Mag.* **12**, 566–576 (1875).

2. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

3. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. vol. 1 (Sinauer Sunderland, MA, 1998).

4. Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* **15**, 246–263 (1886).

5. Zhu, H. & Zhou, X. Statistical methods for SNP heritability estimation and partition: A review. *Comput. Struct. Biotechnol. J.* (2020).

6. Timpson, N. J., Greenwood, C. M., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110 (2018).

7. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

8. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279 (2016).

9. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

10. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).

11. 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).

12. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).

13. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

14. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).

15. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

16. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213 (2013).

17. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).

18. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

19. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

20. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455 (2016).

21. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100**, 15776–15781 (2003).

22. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).

23. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

24. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).

25. Consortium, F. A promoter-level mammalian expression atlas. *Nature* **507**, 462 (2014).

26. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478 (2017).

27. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473 (2012).

28. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

29. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).

30. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

31. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

32. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

33. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

34. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

35. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

36. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

37. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* **51**, 343–353 (2019).

38. de la Torre-Ubieta, L. *et al.* The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* **172**, 289–304 (2018).

39. Liu, L. *et al.* Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 1–11 (2019).

40. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).

41. Soskic, B. *et al.* Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat. Genet.* **51**, 1486–1493 (2019).

42. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).

43. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).

44. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).

45. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).

46. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

47. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

48. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

49. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

50. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).

51. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

52. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

53. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

54. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

55. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

56. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

57. Consortium, Gtex. Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).

58. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).

59. Sieberts, S. K. *et al.* Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci. Data* **7**, 1–11 (2020).

60. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117 (2017).

61. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091 (2015).

62. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

63. van der Wijst, M. G. *et al.* Science Forum: The single-cell eQTLGen consortium. *Elife* **9**, e52155 (2020).

64. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).

65. Benton, M. L., Talipineni, S. C., Kostka, D. & Capra, J. A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *Bmc Genomics* **20**, 1–22 (2019).

66. 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).

67. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).

68. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

69. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

70. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).

71. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265 (2012).

72. 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).

73. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

74. van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).

75. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169 (2019).

76. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).

77. Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, 1–10 (2015).

78. Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* **26**, 238–255 (2016).

79. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

80. Davis, J. E. *et al.* Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. *Cell Syst.* **11**, 75–85 (2020).

81. Jayavelu, N. D., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1–15 (2020).

82. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).

83. Maricque, B. B., Dougherty, J. D. & Cohen, B. A. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res.* **45**, e16–e16 (2017).

84. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* **25**, 713–727 (2019).

85. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci.* **110**, 11952–11957 (2013).

86. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 1–15 (2019).

87. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).

88. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).

89. Cheung, R. *et al.* A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Mol. Cell* **73**, 183–194 (2019).

90. Rabani, M., Pieper, L., Chew, G.-L. & Schier, A. F. A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. *Mol. Cell* **68**, 1083–1094 (2017).

91. Litterman, A. J. *et al.* A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.* **29**, 896–906 (2019).

92. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).

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

94. 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).

95. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

96. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

97. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 1–14 (2017).

98. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* **11**, 1–16 (2020).

99. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).

100. Castaldi, P. J. *et al.* Identification of functional variants in the FAM13A chronic obstructive pulmonary disease genome-wide association study locus by massively parallel reporter assays. *Am. J. Respir. Crit. Care Med.* **199**, 52–61 (2019).

101. Lu, X. *et al.* Global discovery of lupus genetic risk variant allelic enhancer activity. *Nat. Commun.* **12**, 1–13 (2021).

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

103. Matoba, N. *et al.* Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Transl. Psychiatry* **10**, 1–14 (2020).

104. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).

105. Banfi, S. *et al.* Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.* **7**, 513–520 (1994).

106. Tanzi, R. E. & Bertram, L. Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell* **120**, 545–555 (2005).

107. Hernandez, D. G., Reed, X. & Singleton, A. B. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J. Neurochem.* **139**, 59–74 (2016).

108. Schneider, L. A resurrection of aducanumab for Alzheimer's disease. *Lancet Neurol.* **19**, 111–112 (2020).

109. Balendra, R. & Isaacs, A. M. C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nat. Rev. Neurol.* **14**, 544–558 (2018).

110. Neumann, H. & Daly, M. J. Variant TREM2 as risk factor for Alzheimer's disease. *N Engl J Med* **368**, 182–4 (2013).

111. Baker, M. *et al.* Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature* **442**, 916–919 (2006).

112. Sheng, J., Su, L., Xu, Z. & Chen, G. Progranulin polymorphism rs5848 is associated with increased risk of Alzheimer's disease. *Gene* **542**, 141–145 (2014).

113. Lee, J.-M. *et al.* Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell* **162**, 516–526 (2015).

114. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet. Med.* **18**, 421–430 (2016).

115. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

116. Bellenguez, C. *et al.* New insights on the genetic etiology of Alzheimer's and related dementia. *medRxiv* (2020).

117. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

118. Nicolas, A. *et al.* Genome-wide analyses identify KIF5A as a novel ALS gene. *Neuron* **97**, 1268–1283 (2018).

119. Kouri, N. *et al.* Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. *Nat. Commun.* **6**, 1–7 (2015).

120. Ferrari, R. *et al.* Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* **13**, 686–699 (2014).

121. Chia, R. *et al.* Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat. Genet.* **53**, 294–303 (2021).

122. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

123. Gan, L., Cookson, M. R., Petrucelli, L. & La Spada, A. R. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nat. Neurosci.* **21**, 1300–1309 (2018).

124. Yokoyama, J. S. *et al.* Shared genetic risk between corticobasal degeneration, progressive supranuclear palsy, and frontotemporal dementia. *Acta Neuropathol. (Berl.)* **133**, 825–837 (2017).

125. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

126. Efthymiou, A. G. & Goate, A. M. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener.* **12**, 1–12 (2017).

127. Bowles, K. *et al.* 17q21. 31 sub-haplotypes underlying H1-associated risk for Parkinson's disease and progressive supranuclear palsy converge on altered glial regulation. (2019).

CHAPTER 2

Characterization of AD and PSP GWAS loci using MPRA

Abstract

Predicting functionality of noncoding variation is one of the major challenges in modern genetics. I employed massively parallel reporter assays to screen 5,706 noncoding variants from three genome-wide association studies for both Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP), two neurodegenerative disorders representing significant global disease burden. I established the robustness and reproducibility of this approach by finding high intra and inter-replicate correlations of predicted variant effects and by confirming enrichment of functional genomic annotations within transcriptionally active library elements. I subsequently identified 320 functional regulatory polymorphisms (SigVars) comprising putative causal signal at 27 of 34 unique tested loci. These predictions were further refined using functional genomic annotations from relevant brain cell types to identify 55 high-confidence causal variants. SigVars were also found to be enriched within functional annotations enriched within neurons in PSP and microglial in AD, confirming that divergent cell types may mediate genetic risk for these related disorders. Finally, regulatory variants defined by my screen were used to systematically characterize 17q21.31 and 19q13.32, two complex regions harboring extended linkage disequilibrium, and nominate novel candidate risk genes in these loci including *PLEKHM1* in PSP and *APOC1* in AD. Thus, the successful prioritization of noncoding regulatory variation associated with AD and PSP demonstrates the utility of high-throughput experimental approaches in the functional dissection of GWAS loci.

Introduction

Neurodegenerative disorders such as Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP) are a major and growing cause of morbidity and mortality worldwide, with AD alone expected to impact 135 million individuals by 2050 [1]. Given the lack of any disease modifying therapeutics, there is a significant need for investigation of causal disease mechanisms. Sporadic AD and PSP, both known as tauopathies because of the pathological deposition of tau in the brains of affected individuals [2], are complex polygenic traits with heritability estimates of between 40-80% [3,4]. Over the last decade a number of genome-wide association studies (GWAS) have identified numerous susceptibility loci [5–12]. However, most of these loci fall in noncoding regions of the genome and encompass numerous variants due to linkage disequilibrium (LD), which has hampered the identification of underlying regulatory variants and associated risk genes [13–15]. This has posed a particular challenge for the interpretation of extended complex haplotypes harboring multiple independent association signals such as the H1 pan-neurodegenerative risk haplotype at 17q21.31 that includes the *MAPT* locus [12,16,17], and the AD risk locus at 19q13.32 that harbors *APOE* [18–20].

Although a number of statistical fine-mapping approaches have been developed to identify causal GWAS variants, these methods perform poorly on underpowered datasets or regions of extended LD (reviewed in [21]). Similarly, prioritization algorithms that score variant pathogenicity by leveraging features such as evolutionary conservation and chromatin annotations underperform in noncoding regions of the genome, or are nonspecific [22,23] (Chapter 3). It is becoming increasingly recognized that functional methods are necessary to identify true causal variants within most loci, but the sheer numbers of variants challenge most experimental approaches. Massively parallel reporter assays (MPRA) provide a solution, enabling high-

throughput experimental characterization of the transcriptional-regulatory potential of noncoding DNA elements [24–27]. In an MPRA, many regulatory elements are cloned into an expression vector harboring a reporter gene and a unique DNA barcode to create an expression library that is assayed via high-throughput sequencing [24]. MPRAs have prioritized functional common variation for lymphoblastoid eQTLs [28], red blood cell traits [29], cancer [30], adiposity [31], and osteoarthritis [32] (Chapter 1). However, they have not been systematically applied to neurodegeneration or any neurologic disorder.

In this chapter, I describe the design and implementation of an MPRA to characterize common genetic variation associated with two distinct neurodegenerative disorders that both share tau pathology, AD and PSP. I screen 5,706 unique variants comprising 34 genome-wide significant or suggestive loci derived from three GWASs [6,7,10], including a comprehensive assessment of disease-associated variants within the 17q21.31 and 19q13.32 regions. I identify 320 variants with significant allelic skew at a conservative false discovery rate, and find enrichment of these functional variants within genomic features associated with transcriptional regulation.

Materials and Methods

Identification of candidate variants

MPRA Stage 1: I selected all genome wide significant ($p < 5\times10^{-8}$) variants from an AD and PSP GWAS [6,10]. I then identified all variants with a MAF > 5% in LD ($r^2 > 0.8$) with these variants in Europeans (CEU + FIN + GBR + IBS + TSI; 1000 Genomes Phase 3) using the LDProxy tool accessed through the LDlink API. I subsequently filtered out indels, multiallelic, and coding variants. Both alleles of each variant were centered in 162 bp of genomic context (hg19/37)

using the biomaRt (2.44.0) and BSgenome (1.56.0) packages in R (4.0.0) to create oligos. I then removed oligos containing KpnI, MluI, SpeI, and XbaI restriction sites needed for library cloning, leaving 5,223 total variants. Finally, I appended 5' (CTGAGTACTGTATGGGCGACGCGT) and 3' (GGTACCGACAAAAGTGTCAACTGT) PCR adaptor sequences to each oligo and synthesized the library on an Agilent 15k 210-mer array.

MPRA Stage 2: I replicated a selection of 186 variants with significant allelic skew ("SigVars"; FDR adjusted $q < 0.01$) identified in MPRA 1 and 140 negative control variants. For 212 variants I also created oligos: 1) in reverse complement orientation, 2) with the variant located in the bottom third of the genomic context (*i.e.* 121 bp upstream and 40 bp downstream genomic context), and 3) in the reverse orientation (discussed further in Chapter 6). Furthermore, in MPRA 2 I attempted to re-assess variants that dropped out of MPRA 1 (defined below). Finally, I assessed the lead SNPs from additional significant loci from two AD GWAS [6,7], as well as 4 PSP genome-wide suggestive loci [10] . LD partners were identified as above, constituting an additional 483 variants. The final MPRA 2 library was synthesized in duplicate on an Agilent 7.5k 210-mer array. All tested loci for both MPRA stages are summarized below in Table 2-1.

MPRA 1		
GWAS	**#**	**Loci**
Lambert et al., 2013 - Stage 1	14 AD loci	CR1, BIN1, CD2AP, EPHA1, CLU, MS4A6A, PICALM, ABCA7, HLA-DRB1/5, PTK2B, SORL1, SLC24A4/RIN3, DSG2, 19q13.32/APOE
Chen et al., 2018	5 PSP loci	MAPT/17q.21.31, MOBP, STX6, RUNX2, SLC01A2
MPRA 2		
GWAS	**#**	**SNPs**
Lambert et al., 2013 - Stage 3	6 AD SNPs	rs3865444, rs35349669, rs1476679, rs10838725, rs17125944, rs7274581
Kunkle et al., 2019	11 AD SNPs	rs12539172, rs3740688, rs17125924, rs12881735, rs3752246, rs6024870, rs7920721, rs138190086, rs593742, rs7185636, rs2830500
Chen et al., 2018	4 PSP lead SNPs	rs12125383, rs147124286, rs2045091, rs114573015

Table 2-1. Loci (MPRA 1) or lead SNPs (MPRA 2) used to identify common variants tested in this study. Loci refers to a collection of all genome-wide significant SNPs in a given linkage region, annotated to the nearest gene.

Custom MPRA vector design

The pAAV-stop-MCS-bGH plasmid was created as follows: The multiple cloning site (MCS) and bGH-polyA sequence from the Donor_eGP2AP_RC plasmid (Addgene #133784) was cloned into the pAAV.CMV.PI.EGFP.WPRE (Addgene #105530) backbone using NheI and SphI restriction sites. To prevent transcriptional readthrough from the AAV ITR, this vector was re-cut at the NheI site and the transcriptional insulator element from pGL4.23 (Promega, E8411) was inserted using Gibson Assembly [33]. The pMPRAdonor2-eGFP plasmid was created by cloning the eGFP open reading frame into the pMPRAdonor2 plasmid (Addgene #49353) digested with NcoI and XbaI restriction enzymes.

Library construction

Sequences for primers described below are listed in Appendix A – Supplemental Materials and Methods.

Step 1: I amplified and attached 20 bp degenerate barcodes to the oligo library by emulsion PCR (Chimerx, 3600-01). I performed four 50 uL PCR reactions with individual mixtures containing: 2 pmol of library, 1 uM of barcode_new_F primer, 1 uM of barcode_N_R primer, 200 uM dNTPs, 0.25 mg/mL acetyl-BSA (Thermo Fisher Scientific, AM2614), and 2 U of Phusion Hot Start II DNA Polymerase (Thermo Fisher Scientific, F549S) in 1X HF buffer. Thermal cycle conditions were: initial denaturation for 1 min at 95°C, followed by 20 cycles of 10 sec at 95°C, 20 sec at 61°C, and 20 sec at 72°C (2.5°C/sec ramp rate), followed by a final extension for 5 min at 72°C. Emulsions were broken with butanol, pooled, and purified per manufacturer's

instructions on spin columns. The amplified library and the pAAV-stop-MCS-bGH plasmid were digested overnight using SpeI-HF and MluI-HF enzymes and purified using Streptavidin M-270 Dynabeads (Thermo Fisher Scientific, 65305) or gel purification respectively (28704, Qiagen). An 80 uL T7 ligation reaction (NEB, M0318S) containing 200 ng of digested plasmid and 37.7 ng of library was performed followed by cleanup and electroporation into DH5α electrocompetent cells (NEB, C2989K). Transformed bacteria were pooled, serially diluted, and plated overnight at 37ºC. Colonies from the dilution plate containing the number of bacterial colonies approximating 50-fold library coverage were collected and grown, followed by Maxiprep library extraction (Thermo Fisher Scientific, K210016).

Barcode mapping: Sequencing was performed to create a lookup table mapping barcodes to oligos. Oligo-barcode sequences were amplified from 2 ng of plasmid using flanking PCR primers (BC_map_P5_Rev and BCmap_P7_For) that added P5 and P7 adaptor sequences. Amplicons were sequenced by the UCLA TCGB core using an Illumina NextSeq 550 system (PE 2x150 bp) using custom Read 1 (BCmap_R1Seq_Rev) and Read 2 (BCmap_R2Seq_For) sequencing primers. Reads were merged using the BBMerge tool [34] and barcodes filtered and assigned to oligos using a python script. Briefly, reads that did not perfectly match library oligos were discarded. Barcodes represented by fewer than three reads were dropped. Ambiguously mapped barcodes were then filtered as follows: I bootstrapped an empirical distribution of oligo Levenshtein distances (python-Levenshtein 0.12.0) to determine a cutoff score (1st percentile of distances). I then discarded barcodes where any pairwise read distance was greater than this cutoff score. The MPRA_barcode_mapping.py python script is provided: https://github.com/ycooper27/Tauopathy-MPRA.

Step 2: The MinP-eGFP fragment was amplified from the pMPRAdonor2-eGFP vector using Amp_minPLuc2_For and Amp_minPLuc2_Rev primers and both the plasmid library and fragment were sequentially digested using KpnI-HF and XbaI enzymes. These were used in a T7 ligation reaction containing 200 ng plasmid and 125 ng fragment. The ligation product was transformed into DH5α competent cells followed by plasmid isolation. The final library was configured with the 162 bp oligo upstream of the minimal promoter and the 20 bp barcode located in the 3' UTR of the eGFP transcript (Figure 2-1).

Massively Parallel Reporter Assay

MPRA was performed with 6 biological replicates each consisting of ~8 million HEK293T cells. I transfected 10 ug of library plasmid per replicate using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008). 24 hours post-transfection, cells were dissociated, pelleted, and washed with DPBS. Replicates were lysed in 1 mL of RLT buffer and homogenized using QIAshredder columns (Qiagen, 79654). Total RNA was extracted and eluted in 100 uL RNase-free water using an RNeasy Mini Kit (Qiagen, 74104) with on-column DNase I digestion (Qiagen, 79254). I then extracted mRNA from 75 ug total RNA per sample using a Dynabeads mRNA Purification Kit (Thermo Fisher Scientific, 61006). Residual DNA was removed from 1 ug of mRNA using ezDNase (Thermo Fisher Scientific, 11766051) followed by reverse transcription using a custom primer (Lib_Hand_RT) and the SuperScript IV First Strand Synthesis Kit (Thermo Fisher Scientific, 18091050). RT was performed for 80 min at 52°C. I then amplified 10 uL of cDNA in 100 uL PCR reactions using NEBNext Ultra II Q5 Master Mix (NEB, M0544S) and Lib_Seq_eGFP_F2 and Lib_Hand primers for either 8 or 3 cycles (MPRA

1 and 2 respectively). Likewise, 5 replicates of 25 ng plasmid DNA (MPRA 1) or 4 replicates of 5 ng plasmid (MPRA 2) were amplified in 50 uL reactions using Lib_Seq_eGFP_F2 and Lib_Hand_RT primers. PCR products were purified using 0.6X-1.2X KAPA Pure Beads (Roche, KK8000) and then further amplified using P5_seq_eGFP_F2 and P7_Ind_#_Han primers for 7 PCR cycles to add P5, P7, and unique 8 bp Illumina index sequences. Following SPRI cleanup, amplicons were sequenced using an Illumina NextSeq 550 (1x20) or NovaSeq 6000 SP flow cell (SR 1x26 cycles) with 5% PhiX spike-in and custom Read 1 (Exp_eGFP_Seq_F2) and Index (Exp_Ind_Seq_P) primers.

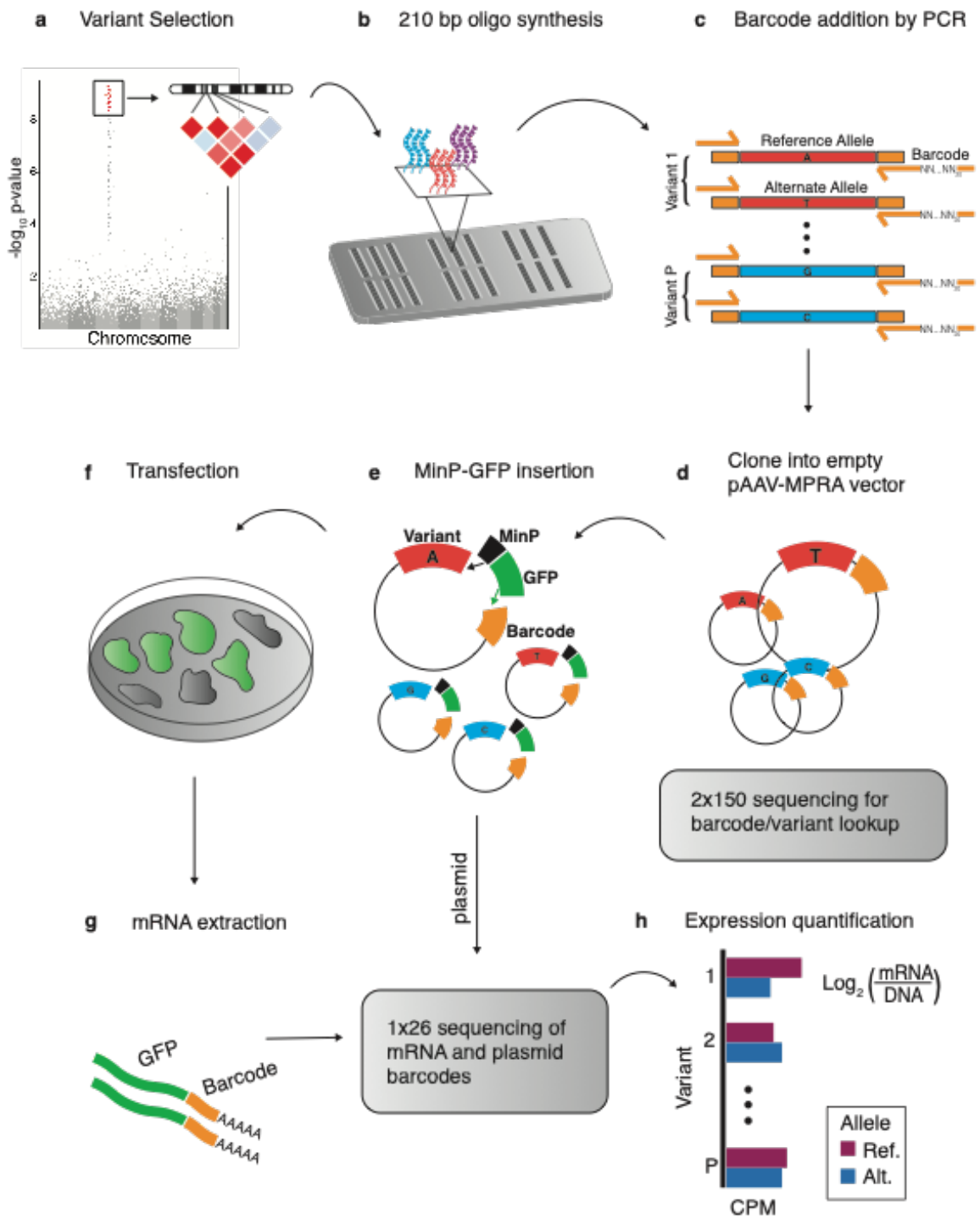


Figure 2-1. Technical schematic for massively parallel reporter assay (MPRA). (**A**) I selected all bi-allelic noncoding genome-wide significant ($p > 5\times10^{-8}$) or locus lead SNPs and LD partners ($r^2 > 0.8$) from three GWAS. (**B**) The reference and alternate alleles of each variant were centered within 162 bp of genomic context (hg19) and synthesized on an Agilent array. (**C**) Oligos were amplified by PCR using primers complementary to flanking shared adaptor sequences. This step also attached 20 nucleotide random barcodes and restriction enzyme sites.

(**D**) Oligos were digested and cloned into an expression library, followed by paired-end 2x150 bp sequencing to map barcodes to associated oligos. (**E**) A minimal promoter and eGFP gene was inserted between the oligo and barcode, with the oligo upstream of the minimal promoter and the barcode in the 3'UTR of eGFP. (**F**) The expression library was transfected into HEK293T cells for 24 hours. (**G**) following mRNA extraction, mRNA and plasmid barcodes were amplified, pooled, and sequenced using indexed single-end sequencing. (**H**) mRNA barcode counts were normalized to DNA counts, and the median normalized barcode count was taken as an activity summary score for each allele.

MPRA data analysis

Preprocessing: Raw reads were trimmed to only contain the 20 bp barcode and aligned to the previously generated oligo-barcode lookup table to create a barcode count matrix. Barcode reads that did not perfectly match were discarded. The MPRA_BC_counter.py mapping script is provided: https://github.com/ycooper27/Tauopathy-MPRA.

Barcodes were filtered such that at least 1 count had to be detected in every RNA replicate and at least 5 counts detected in each DNA replicate (to ensure stability of the log-ratios). A pseudocount of 1 was added to each barcode count, which was then normalized to sequencing depth to create counts per million reads per replicate (CPM). I then computed $\log_2$ RNA/DNA ratios for each barcode (DNA = median count across plasmid replicates), which were then quantile normalized between replicates using the preprocessCore (1.50.0) package. Variants with fewer than 5 unique barcodes for either allele were removed from further analysis.

Oligo activity measurements: To calculate transcriptional activity scores for each 162 bp oligo, allele-level summary statistics were computed as the median of the $\log_2$ barcode ratios. This value was then averaged between reference and alternate alleles to create an activity score for each oligo. To determine significance, this activity score was compared to the median activity value for the entire library using a one-sample Mann-Whitney-U test (two-tailed; n = 6

replicates), which was subsequently adjusted for multiple comparisons. “Active” elements for subsequent analyses had an increased RNA/DNA ratio compared with the library median at a Bonferonni adjusted $p < 0.05$ (or FDR adjusted $q < 0.01$ where described; Benjamini Hochberg method).

SigVar calculations: To identify variants with significant allelic skew (SigVars), $\log_2$ barcode ratios were combined across all 6 replicates by taking the median value. For each variant, a two-way Mann-Whitney-U test comparing barcode counts between each allele was used to identify allelic skew. SigVars were defined at an FDR threshold $q < 0.01$ (FDR adjustment, Benjamini-Hochberg method; further discussed below in Statistical Reporting). MPRA $\log_2$ effect sizes were defined as the median summed normalized barcode count for the alternate allele - reference allele.

Quality control: Intra-experiment barcode reproducibility was defined as the mean pairwise correlation of each normalized barcode (RNA/DNA) count across all technical replicates. Allele correlation was determined by first finding the median normalized barcode count for each allele followed by determining mean pairwise correlation across all technical replicates (Both Pearson’s r and Spearman’s rho computed). Between experiment correlations for reference allele activity scores and variant effect sizes were also determined for 326 variants replicated in MPRA 2 estimate inter-experiment reproducibility.

Bioinformatic analyses

Jaccard Index calculations: I download DNase I hotspot files from the ENCODE project server (https://www.encodeproject.org/) [35] for brain cell types and tissues (cell types, accessions listed in Appendix A – Supplemental Materials and Methods. Pairwise Jaccard indices between all samples were calculated using the Jaccard tool from BEDTools (2.29.2) [36] and plotted as a heatmap. Biological replicates (rep #2) were then discarded to avoid artificial inflation (leaving only one sample per cell type), and the average pairwise Jaccard index for each cell type with all other cell types was computed. GWAS overlap was calculated for a given cell type by first intersecting all tested GWAS variants with that cell type's DNase I peaks. Then, only the DNase-overlapping variants were intersected with DNase I peaks from the other cell types (pairwise) and the mean proportion shared was computed.

Chromatin annotation enrichment: I downloaded narrowPeak files for HEK293 DNase-seq, histone ChIP-Seq, and TF-ChIP-seq marks (accessions listed in Appendix A – Supplemental Materials and Methods) from the ENCODE project server [35]. I then determined overlap between these marks and MPRA "active" and "repressive" elements using the GenomicRanges R package (1.40.0) assuming a minimum of 1 bp overlap between the 162 bp oligo and the chromatin mark. Enrichment was calculated for active or repressive elements against a background set of all other tested oligos using a Fisher's exact test, with $\log_2$ odds ratios and 95% confidence intervals reported.

SigVar functional annotations: SigVars from this study were annotated for TFBS disruption and overlapped with functional brain annotations (Supplemental Table 1). I calculated TFBS

disruption using both the motifbreakR package using the HOCOMOCO v10 TF binding model (filtered for a binding threshold of $p < 1\times10^{-4}$ and "strong" predicted effects) as well as the SNPS2TFBS webtool [37]. Additionally, published enhancer and promoter annotations for sorted microglia, neurons, astrocytes, and oligodendrocytes were downloaded and converted to bed files using the ucsc-bigbedtobed tool and overlapped with SigVars using BEDTools intersect [36,38]. I also identified SigVars overlapping high-confidence multi-tissue enhancers defined by the HACER database [39]. Promoter/Enhancer accessions are provided in Appendix A – Supplemental Materials and Methods.

LD clustering: To calculate LD between SigVars within the 17q21.31 locus, I downloaded chr17 VCF files from the 1000 Genomes FTP server. The VCF was subsetted for 90 unrelated individuals of CEU ancestry and reformatted to get cumulative allele frequencies using PLINK 1.9 [40]. Genotype files were loaded into R and used to create LD clusters for SigVars within 17q21.31 using the clqd function (CLQDmode = "maximal") from the gpart R package (1.6.0) [41].

Cell culture

I obtained HEK293T (CRL-3216) cells from ATCC. HEK293T cells were cultured in DMEM containing GlutaMAX (Thermo Fisher Scientific, 10566016) supplemented with 10% FBS and 1% Sodium Pyruvate (11360070).

Statistical reporting

Statistical analysis was performed using the stats package in R. All hypothesis testing was two-sided. Unless otherwise stated, all enrichment analysis was performed using a Fisher's exact

test. MPRA allelic skew multiple testing correction was performed as follows: Variants from MPRA 1 were combined with additional unique variants tested in MPRA 2 (total 5340 variants) and Mann-Whitney-U p-values were FDR-adjusted (BH method). SigVars were called at a threshold of $q < 0.01$. To assess variant reproducibility, variants replicated, *i.e.* re-tested (320 total) in MPRA 2 were considered separately, and assigned significance at a Bonferroni-adjusted $q < 0.05$.

Data visualization

Variant genomic annotations were determined and plotted using the annotatr (1.14.0) Bioconductor package and ggplot2 from the tidyverse collection (1.3.0). Heatmaps were generated using the pheatmap R package (1.0.12). The circle Manhatten plot was created using the CMplot R package (https://github.com/YinLiLin/R-CMplot). 17q21.31 LD plots were created using the BigLD function from the gpart package [41]. All other data were visualized using ggplot2 with the GGalley package extension.

Results

MPRA to identify candidate regulatory GWAS variants

I conducted a staged analysis to identify regulatory variants underlying GWAS loci for two neurodegenerative tauopathies – Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP) – using massively parallel reporter assays (MPRA) (Figure 2-2). In stage 1 (MPRA 1), I identified all variants in linkage disequilibrium (LD; $r^2 > 0.8$) with the 1,090 genome-wide significant ($p < 5\times10^{-8}$) variants from an AD GWAS [6] and 3,626 genome-wide significant variants from a PSP GWAS [10]. After filtering for bi-allelic noncoding variants, this resulted in a

list of 5,223 variants encompassing 14 AD and 5 PSP GWAS loci. Both alleles of each variant were centered in 162 base pairs (bp) of genomic context, synthesized as 210 bp oligonucleotides, and cloned into a custom MPRA vector along with degenerate barcodes to create an expression library (Figure 2-1; Methods). In stage 2 (MPRA 2), I sought to replicate 326 variants screened in MPRA 1 to asses reproducibility, test the importance of oligo configuration on assay performance, and screen an additional 483 variants encompassing 11 new loci from 2 recent AD GWAS [6,7] and 4 new suggestive loci for PSP [10] (Figure 2-2; Table 2-2; Methods).

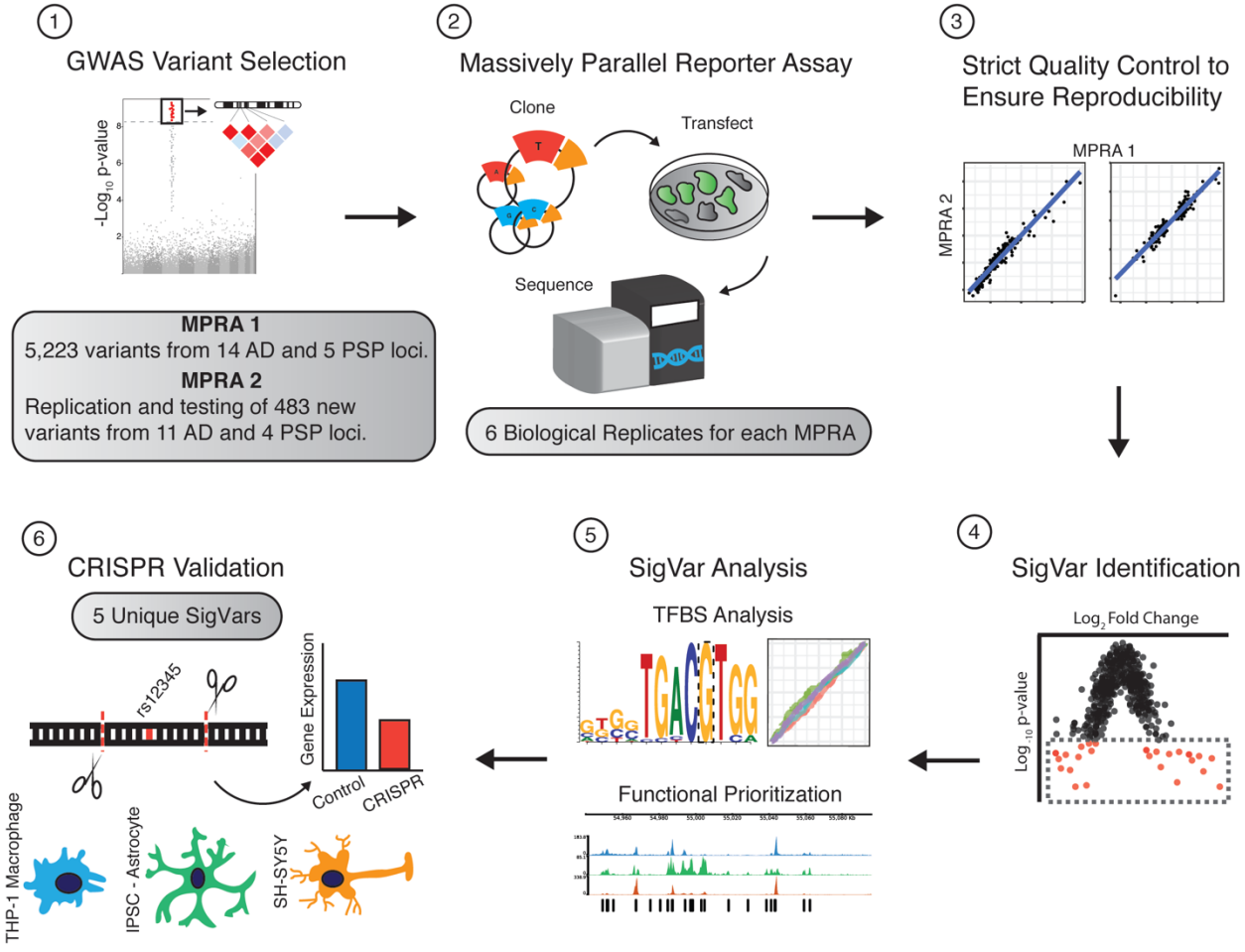


Figure 2-2. Project workflow**:** 1) 5,223 genome-wide significant variants and LD partners encompassing 14 AD and 5 PSP GWAS loci were selected in MPRA 1. For MPRA 2, select variants identified in MPRA 1 were replicated. An additional 483 variants from 11 AD and 4 PSP loci were also tested. 2) Both alleles of each variant were barcoded and cloned into an

expression library that was transfected into HEK293T cells. Allele expression was quantified by next-generation sequencing of associated barcodes. 3) Strict quality control was performed to confirm within-assay and between-experiment reproducibility. 4) Comparison of expression between alleles enabled identification of variants with significant allele-specific transcriptional skew (SigVars). 5) SigVars were used for benchmarking of computational prediction algorithms (see Chapter 3) and transcription factor binding site-disruption analysis (see Chapter 4). SigVars were further prioritized using brain-specific genomic annotations and 6) five top variants were selected for validation using CRISPR/Cas9 genome-editing in brain-relevant cell lines (THP-1, IPSC-astrocytes, SH-SY5Y; see Chapter 5).

Chr	Pos/Cytoband	Lead SNP	Annotated Gene	GWAS[#]	MPRA Stage	# SNPs Tested	SigVars
1	85603051	rs114573015	*WDR63*	PSP	2	4	0
1	180993146	rs57113693	*STX6*	PSP	1	44	2
1	207692049	rs6656401	*CR1*	AD 1	1	22	2
1	221995092	rs12125383	*DUSP10*	PSP	2	51	3
2	127892810	rs6733839	*BIN1*	AD 1	1	55	7
2	234068476	rs35349669	*INPP5D*	AD 1	2	45	2
3	39510287	rs10675541	*MOBP*	PSP	1	80	2
6	32578530	rs9271192	*HLA-DRB5– HLA-DRB1*	AD 1	1	445	13
6	45499614	rs35740963	*RUNX2*	PSP	1	65	1
6	47487762	rs10948363	*CD2AP*	AD 1	1	71	3
7	100004446; 100091795	rs1476679; rs12539172	*ZCWPW1; NYAP1*	AD 1&2	2	7	0
7	143110762	rs11771145	*EPHA1*	AD 1	1	29	3
8	27195121	rs28834970	*PTK2B*	AD 1	1	8	2
8	27467686	rs9331896	*CLU*	AD 1	1	11	1
8	131075859	rs2045091	*ASAP1*	PSP	2	49	2
10	11720308	rs7920721	*ECHDC3*	AD 2	2	8	1
11	47557871; 47380340	rs10838725; rs3740688 *	*CELF1/ SPI1-PU.1*	AD 1&2	2	36	5
11	59923508	rs983392	*MS4A6A*	AD 1	1	177	4
11	85867875	rs10792832	*PICALM*	AD 1	1	64	0
11	121435587	rs11218343	*SORL1*	AD 1	1	3	0
12	21314281	rs7966334	SLC01A2	PSP	1	8	0
12	53788003	rs147124286	*SP1*	PSP	2	79	5
14	53400629; 53391680	rs17125944; rs17125924	*FERMT2*	AD 1&2	2	26	2
14	92926952; 92932828	rs10498633; rs12881735	*SLC24A4/RIN3*	AD 1&2	1	4	1
15	59045774	rs593742	*ADAM10*	AD 2	2	14	3
16	19808163	rs7185636	*IQCK*	AD 2	2	130	17

17	17q21.31	NA	*MAPT*	PSP	1	3482	194
17	61538148	rs138190086	*ACE*	AD 2	2	6	1
18	29088958	rs8093731	*DSG2*	AD 1	1	11	2
19	1063443	rs4147929	*ABCA7*	AD 1	1	8	0
19	19q13.32	NA	*APOE/TOMM40*	AD 1	1	640	37
19	51727962	rs3865444	*CD33*	AD 1	2	6	0
20	55018260; 54997568	rs7274581; rs6024870	*CASS4*	AD 1&2	2	17	4
21	28156856	rs2830500	*ADAMTS1*	AD 2	2	1	1

Source GWAS for tested loci and variants. PSP = Chen *et al.,* 2018 ; AD 1 = Lambert *et al.*, 2013; AD 2 = Kunkle *et al.,* 2019.
* Locus re-mapped in AD GWAS 2. Lead SNPs are not in LD. Both lead SNPs were tested here.

Table 2-2. Description of GWAS loci and variation tested in this study. Shown are the locus lead SNPs and annotated genes as described in the listed GWASs. For each locus: the MPRA stage in which it was tested, the number of variants tested per locus, and the number of variants with significant allelic skew (FDR q < 0.01; SigVars) ultimately identified.

The performance of MPRAs to detect allelic skew depends upon high library transfection efficiency [42], necessitating the use of easy to transfect cell lines, which in published studies have included HEK293T, K562, and HepG2 cells, among others [28,43]. However, AD and PSP disease risk variants fall within open chromatin across several different neuronal and glial cell types, many of which are non-overlapping [7]. Using available ENCODE DHS data [44], I found poor DNA accessibility overlap between divergent brain cell types, such as astrocytes and neural progenitors (mean Jaccard index = 0.14). Notably, HEK293T cells had the highest mean pairwise Jaccard index (0.22; Figure 2-3a) compared with all brain cell types and tissues, which was not driven by a larger number, or increased width of peaks in HEK293T cells. Moreover, I found that AD and PSP GWAS variants that fell within open chromatin in any brain cell type were also likely to fall within open chromatin in HEK293T cells (mean = 60%; Figure 2-3b). These data indicated that HEK293T cells would provide an optimal model for such high-throughput screening in a single cell line, and they were chosen for the MPRA.

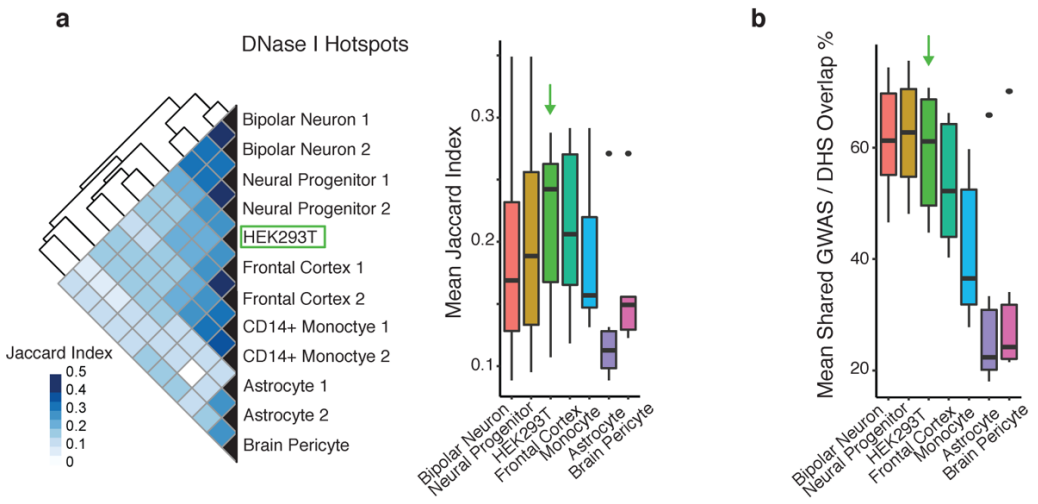


Figure 2-3. comparison of the chromatin landscape between divergent brain cell types. (**A**) left – Heatmap quantifying overlap (Jaccard index) of DNase hotspots (ENCODE project) between primary brain cell types and HEK293T cells. Right – boxplot displaying the mean pairwise Jaccard index for each cell type. (**B**), GWAS variants tested in this study were overlapped with DNase hotspots from each cell type. The boxplot displays the mean pairwise probability that a GWAS variant within open chromatin in a given cell type also overlaps open chromatin in another cell type. Green arrows highlight HEK293T. Error bars = S.E.M.

I performed MPRA (n = 6 biological replicates), obtaining activity measurements from at least 5 unique barcodes for both alleles for 4,732 of 5,223 variants (~91%). Genomic annotations for the variants tested in MPRA 1 are shown in Figure 2-4. Overall, the library achieved a median barcode complexity of 40 barcodes/allele (Figure 2-5a). Intra-library reproducibility was high; allele correlation between replicates had mean Pearson's r = 0.95 (Figure 2-5d) and normalized barcode correlations were consistent with expectations (r = 0.55; Figure 2-5c). Overall, I observed

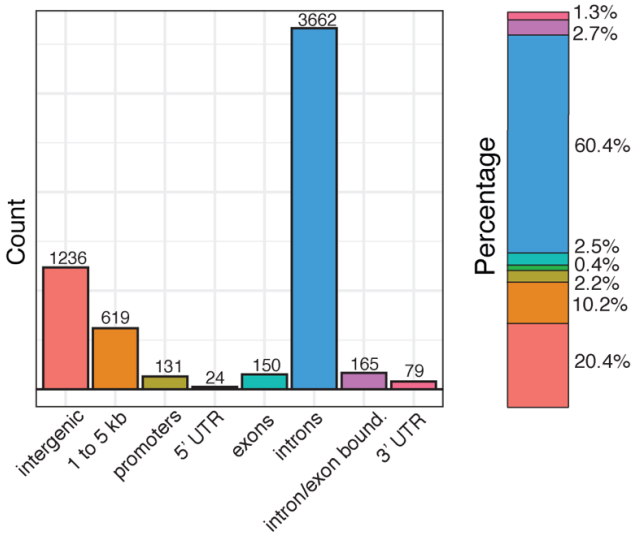


Figure 2-4. Genomic annotations for 5,223 variants tested in MPRA 1.

that ~19% of library elements were transcriptionally active in the assay (Figure 2-6a), in concordance with previous estimates [28,29].

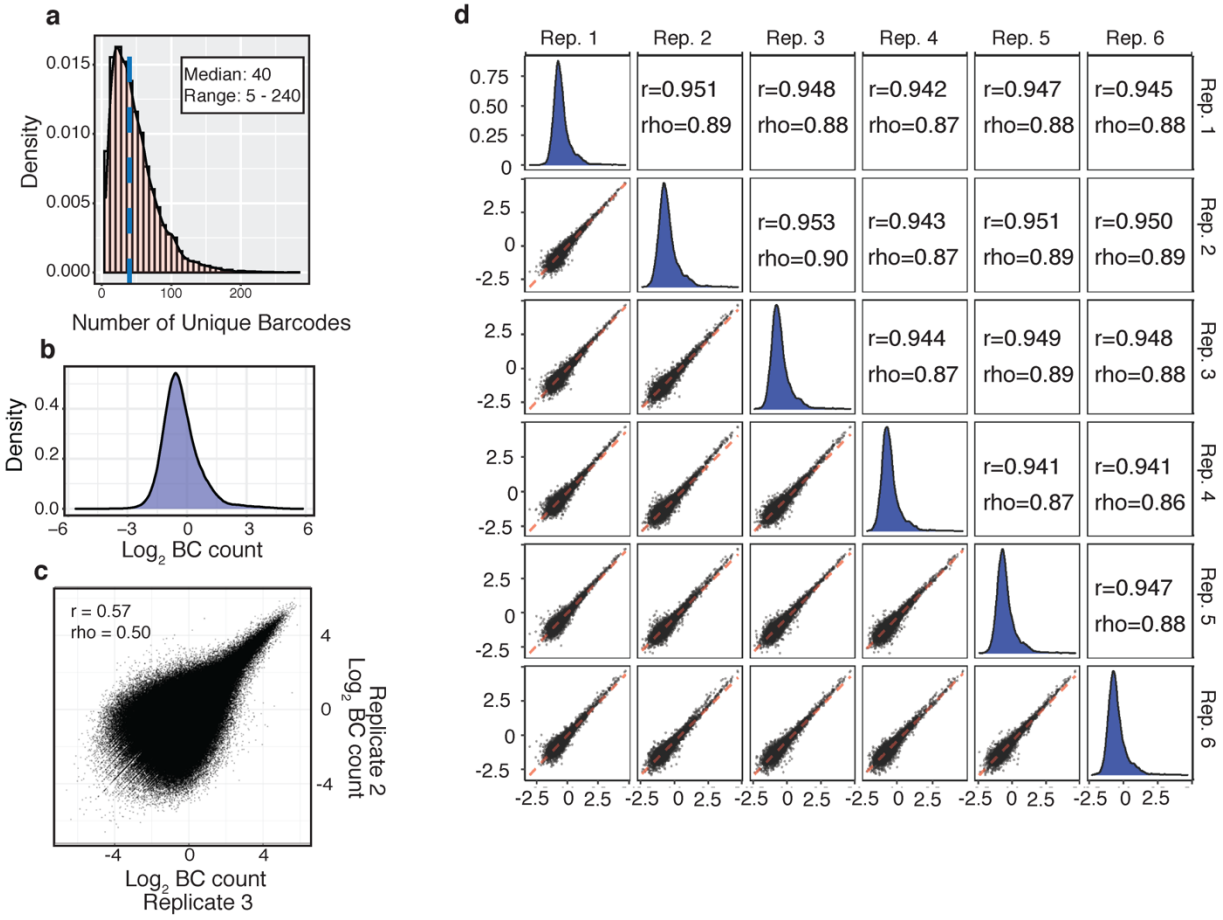


Figure 2-5. MPRA 1 quality control metrics. (**A**) Density plot shows the distribution of uniquely mapped 20 nucleotide barcodes per individual allele (i.e. unique barcodes/allele; 9,464 total alleles) for the full MPRA 1 library, blue line = median (40 barcodes/allele). (**B**) Density plot of $\log_2$ normalized barcode (BC) ratios. (**C**) Representative plot showing the correlation of $\log_2$ normalized barcode counts between technical replicates (Pearson's r = 0.57; Spearman's rho = 0.50; both $p < 2\times10^{-16}$). (**D**) MPRA 1 exhibits high reproducibility between biological replicates (mean r = 0.95; n = 6). Panels show inter-replicate, pairwise correlations of median $\log_2$ normalized barcode counts for all alleles passing filter (n = 9,464). Red line = regression line of best fit, Pearson's correlation, all $p < 2\times10^{-16}$.

I next assessed the functional genome annotations associated with active versus non-active elements. Using available ENCODE data for HEK293T cells, I found depletion of the H3K36me3 mark in active elements (Fisher's exact test; $\log_2$ OR = -0.54, FDR-adjusted q =

0.01). Conversely, active elements were highly enriched for DHS sites ($\log_2$ OR = 1.52, q = 1×10^{-10}), which are indicative of accessible chromatin. Furthermore, H3K27ac and H3K4me3 marks, delineating active enhancers and promoters respectively, were likewise enriched in active elements (Figure 2-6b). ChIP-seq peaks for specific Transcription Factors (TF) were similarly enriched in active elements (Figure 2-6b). I also assessed for enrichment of specific TFBSs within active elements and identified significant enrichments of *SP*/*KLF*, *ETS*, and *AP-1* family members (Figure 2-6c).

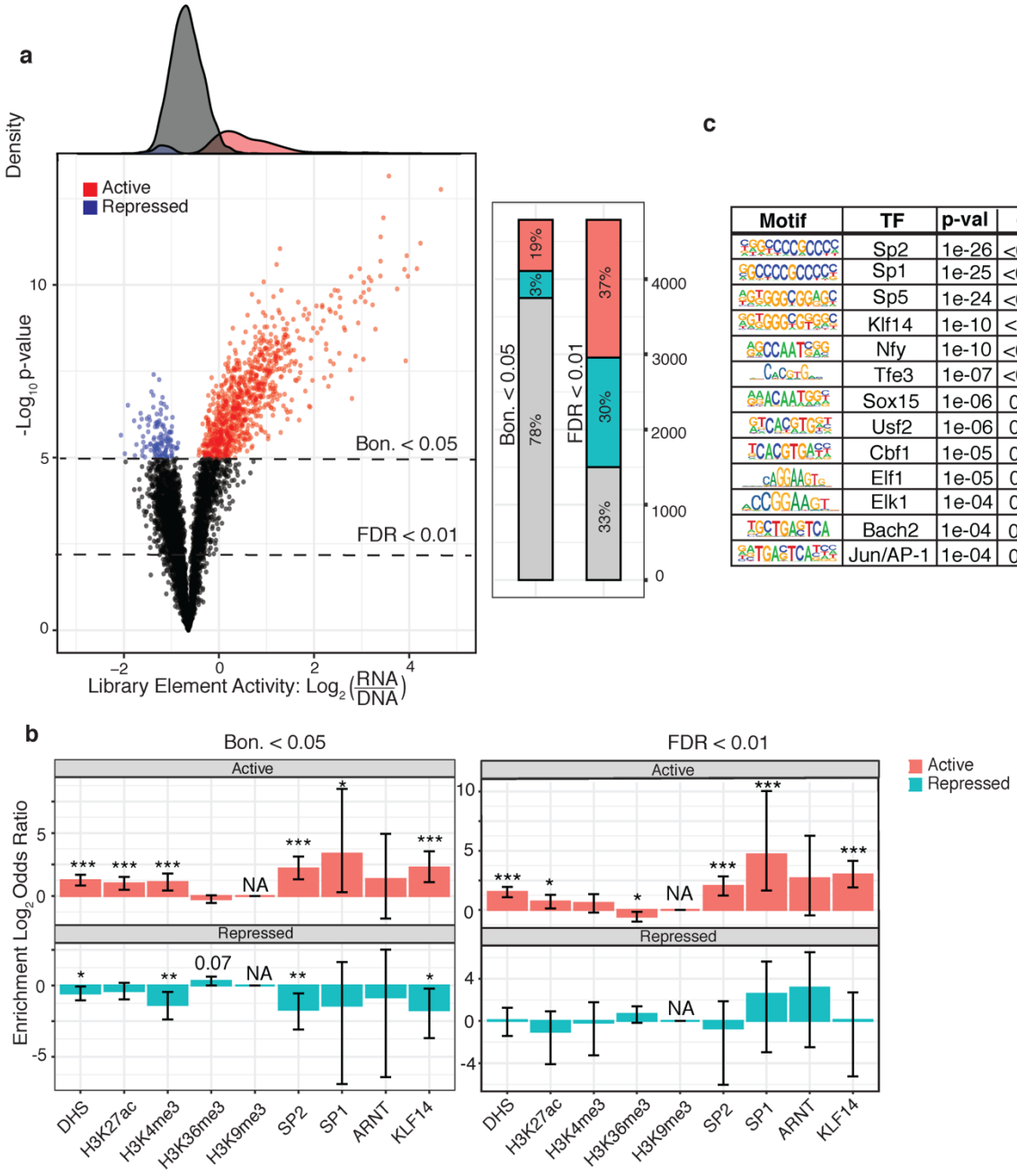


Figure 2-6. Enrichment of active and repressed elements in MPRA 1 with functional genomic features. (**A**) Identification of active and repressed MPRA elements. MPRA library transcriptional activity was quantified by comparing the median normalized barcode count for each element against the median of the whole library (n = 6 replicates; one-sample Mann-Whitney-U test; two-sided, Bonferroni correction). Significantly (Bonferroni $p < 0.05$) increased (active) and decreased (repressed) library elements highlighted on the volcano plot in red and blue respectively. (**B-C**) Active elements are enriched for relevant chromatin features and TFBSs. (**B**) Enrichment $\log_2$ odds ratios (Fisher's exact test) of active and repressed elements (thresholds defined at: left - Bonferroni $p < 0.05$, right – FDR $q < 0.01$) within HEK293T ChIP-seq peaks for both histone and TF marks. Error bars = 95% CI, *** FDR-adjusted $q < 0.001$ (BH

method), * $q < 0.05$. (**C**) Active elements enrich for predicted TF binding motifs using HOMER, including *SP*/*KLF*, and *FOS/JUN* family TFs.

I applied a stringent statistical threshold to identify variants with significantly different transcriptional efficacy between alleles, termed *SigVars* (two-sided Mann-Whitney-U test, FDR $q < 0.01$, Benjamini-Hochberg method), identifying 267 SigVars. I next analyzed the second MPRA (n = 6 biological replicates), which maintained high quality with median library barcode complexity of 54 and high allele level (mean r = 0.99) and barcode level (mean r = 0.84) correlations between replicates (Figure 2-7). Assessments of reproducibility between the separate MPRA experiments also revealed that the assay was robust; the correlations between activity scores (Pearson's r = 0.98, $p < 2\times10^{-16}$) as well as effect sizes (r = 0.94, $p < 2\times10^{-16}$; Figure 2-8a) for replicated variants were both high, and 152 of 186 (82%) re-tested SigVars were reproduced in the second MPRA (replication Bonferroni $p < 0.05$). Placing oligos in the reverse orientation completely abolished reporter activity as expected (Chapter 6).

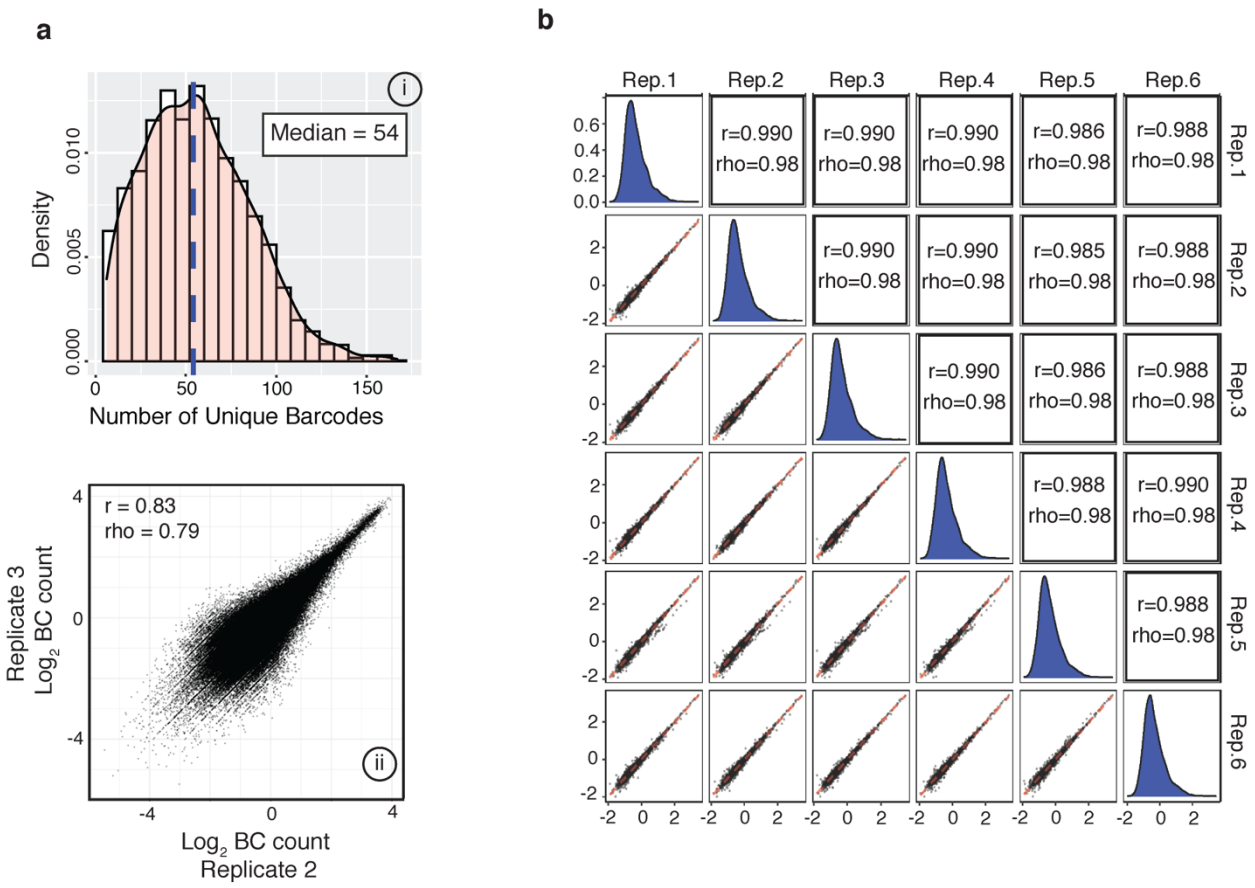


Figure 2-7. MPRA 2 quality metrics. (**A**) Barcode distribution for the 3,072 unique alleles tested in the full MPRA 2 library (blue line = median) with ii) representative plot showing correlation of normalized barcode (BC) counts between technical replicates (Pearson's r = 0.83, Spearman's rho = .79, $p < 2\times10^{-16}$). (**B**) Panels show pairwise correlation of median $\log_2$ normalized barcode counts for all alleles passing filter (n = 3,072) between 6 technical replicates. Red line = OLS regression line of best fit, all $p < 2\times10^{-16}$.

Given the high correspondence between both MPRA experiments, I combined them, obtaining activity measurements from 5,340 of 5,706 (93.6%) assayed variants and identifying 320 unique SigVars distributed across 17 chromosomes (6.0%; Figure 2-8b). I identified SigVars in 27 of 34 (79%) tested GWAS loci with a median of 2 SigVars per locus (Figure 2-8c; Table 2-2). As expected, effect sizes (alt/ref allele) were generally modest (Figure 2-8d), with mean absolute SigVar $\log_2$ fold change of 0.53, consistent with prior work [29]. SigVars were highly enriched within library elements that were transcriptionally active (Figure 2-7a) in this screen

(OR = 6.2; $p < 2\times10^{-16}$; Methods) and were also significantly enriched within DHSs within major brain cell types including astrocytes and neural progenitors, as well as monocytes, which are from the same lineage as microglia in the brain (Figure 2-8e). However, when I separated SigVars derived from AD vs. PSP GWAS loci and identified those that were found to overlap enhancer marks from human brain neurons, microglia, astrocytes, and oligodendrocytes, I saw that the cell types impacted by each disorder were distinct (Methods) [38]. A plurality of AD SigVars fell within microglial enhancers. In contrast, a plurality of PSP SigVars fell within neuronal enhancers, in concordance with recent estimates of cell type specific enrichment in SNP-based heritability for these two disorders [45] (Figure 2-8f). Interestingly, combined across both disorders, 55/78 (71%) of these variants overlapped with cell-type specific enhancer annotations (Figure 2-8g).

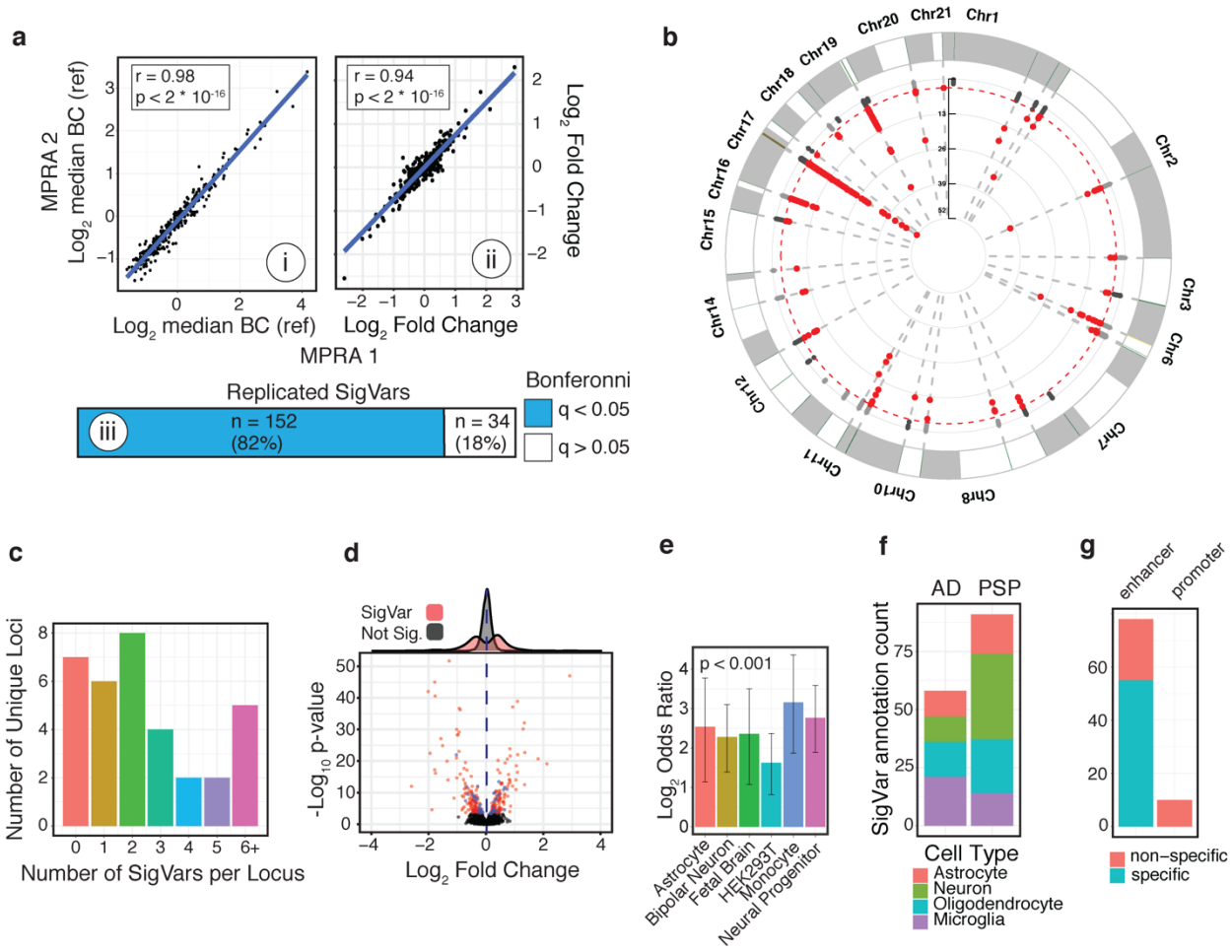


Figure 2-8. Identification of variants with significant allelic skew (SigVars) from both MPRA stages. (**A**) Reproducibility of MPRA across experimental stages. 326 variants, including 186 SigVars identified from the preliminary MPRA (1) were re-tested in a follow-up MPRA (2). i) Reference allele transcriptional activity and ii) $\log_2$ effect sizes (alt/ref allele) show strong correlation (Pearson's r = 0.98, 0.94, $p < 2\times10^{-16}$) between experiments. BC = barcode count. iii) 152 of 186 assessed SigVars from MPRA 1 were replicated in MPRA 2 at replication Bonferroni $p < 0.05$. (**B**) Manhattan plot of 5,340 unique variants successfully tested across both experiments. Red indicates SigVars at FDR-adjusted $q < 0.01$ (BH method). (**C**) Histogram of the number of SigVars identified per GWAS locus LD block (median = 2). (**D**) Volcano plot shows $\log_2$ allelic skew effect sizes and $-\log_{10}$ p-values for 5,340 unique variants tested by MPRA. SigVars for MPRA stage 1 (red) and 2 (blue) highlighted. blue line = median effect size. (**E**) Enrichment $\log_2$ odds ratios (Fisher's exact test) of SigVars within DHSs of various brain cell types (ENCODE project). Error bars = 95% CI, All FDR-adjusted $q < 0.001$. (**F-G**) SigVars were separated into those derived from AD or PSP GWAS and annotated for overlap with enhancer marks from sorted brain tissue (annotations: Nott *et al.,* 2019). Bar plot shows cell-type enhancer annotation counts for overlapping SigVars separated by disease. A plurality of AD SigVars fell within microglial enhancers, while PSP SigVars fell within neuronal enhancers. (**G**) most SigVars overlap enhancers present in only one cell type.

Refinement of SigVar annotations for high confidence predictions of causal variants

I next annotated the functional variants identified in this screen – representing likely causal variants – into those that fell within high confidence promoters or enhancers in neurons, astrocytes, microglia or oligodendrocytes (annotations [38]), as well as those variants that strongly disrupted predicted TFBSs (union of two algorithms [37,46]; Supplemental Table 1). Of 320 SigVars, 233 (73%) had at least one additional functional annotation: 200 (63%) significantly altered TF-binding, 88 (28%) fell within a promoter or enhancer in at least one brain cell type, and 55 (17%) were double annotated for TFBS disruption and an additional promoter/enhancer mark. These 55 SNPs, disrupting TFBSs within annotated promoters or enhancers, represent the highest confidence causal functional variants within ten distinct GWAS loci, including *BIN1*, *HLA-DRB1/5*, *PTK2B*, *CLU*, *ICQK/KNOP1*, 17q21.31/*MAPT*, 19q13.32/*APOE*, *ASAP1*, *SPI1/CELF*, and *CASS4*.

Systematic characterization of complex haplotypes 17q21.31 and 19q13.32

It is particularly challenging to identify causal variants within genomic loci harboring extensive LD using statistical fine-mapping (see Chapter 3), and characterization of these regions can substantially benefit from functional approaches. Considerable common genetic risk for Alzheimer's disease and Progressive Supranuclear Palsy segregate to three such loci. This includes the 17q21.31 locus which harbors *MAPT*, the 19q13.32 locus that harbors *APOE*, and the extended *HLA* type II region on chromosome 6. In the following subsection, I will describe the systematic characterization of regulatory variants within two of these regions, 17q21.31 and 19q13.31, using my MPRA data.

17q21.31: The chromosome 17q21.31 locus is noteworthy for harboring the tau-encoding *MAPT* gene within a common 900 kb inversion-polymorphism [16], and is a major risk locus for PSP (H1 haplotype OR = 4-5) as well as AD, Parkinson's disease (PD), and Corticobasal Degeneration (CBD) [12]. 17q21.31 contains complex haplotypic sub-structural variation and extensive LD, hampering interrogation with traditional statistical genetics approaches. I leveraged the ability to functionally dissect this region, testing 3,482 variants within 17q21.31 in strong LD with lead SNPs from a PSP GWAS [10], comprising approximately ~24% of the more than 14,000 common variants in the region. Of these, I identified a total of 194 SigVars, of which 111 were stringently replicated in both MPRA experiments. Of these replicated SigVars, 20 variants were also double annotated for active chromatin features and TFBS disruption, making them very high confidence causal regulatory variants (Supplemental Table 1).

I next clustered these SigVars based on LD (BigLD algorithm)[41], which identified seven distinct LD clusters within four contiguous LD blocks (Figure 2-9a), suggesting distinct loci within this region. The largest LD cluster includes 42 SigVars within *MAPT* itself (-5 kb upstream of TSS to 3' UTR) highlighting its striking regulatory complexity (Figure 2-9b-c). Of these, 23 were replicated SigVars, 13 of which are variants within annotated enhancers also predicted to disrupt TFBSs (Figure 2-9b). The *MAPT* promoter region overlaps a large CpG island and a number of repetitive transposon elements that may impact gene expression [47]. This region has been previously characterized using serial deletion assays in a variety of cell types [48–50], which identified a core promoter beginning -300 bp from the TSS [47]. In this study, the region -226/-63 (assay ID = 1447) was the 11th most transcriptionally active library element assayed overall, while -349/-186 (ID = 1446) had only modest expression, suggesting a more restricted core promoter starting at -186 upstream of the *MAPT* TSS (Figure 2-9d). I also identified

SigVars within the broader promoter region (-4364/+3292; Figure 2-9c) including rs17770296, which falls in the distal promoter (-2612) and overlaps a MLT1I (ERVL-MALR family) transposable element. Another variant, rs76324150 (+1485), falls within neuronal H3K4me3 peaks and is predicted to disrupt binding of the TF *ZFX* (Figure 2-9d).

SP1 is known to bind the *MAPT* core promoter and regulate Tau expression [51], and was also identified as a suggestive PSP risk locus [10]. Unfortunately, four potentially interesting variants within the proximal promoter region (-144/+1485) dropped out of our assay likely due to the high GC content of the region [47] (Chapter 6). Nevertheless, I identified four SigVars within the *MAPT* gene region predicted to disrupt binding of SP1 (Supplemental Table 1), including rs76839282 which lies within H3K27ac peaks within the long regulatory intron 1 of *MAPT* (Figure 2-9c).

I also identified SigVars within other LD blocks that are predicted to regulate independent risk genes. I highlight rs111392251, a high-confidence regulatory variant located in the promoter of *PLEKHM1* that is predicted to disrupt binding of *IRF*-family TFs (Figure 2-9e-g). *PLEKHM1* regulates autophagosome-lysosome formation [52] and has been previously suggested as a PSP risk gene [9], but has yet to be extensively characterized in a disease context. Other independent risk genes in distinct LD blocks implicated here include *MAP3K14* and *LRRC37A4P* (Table 2-3).

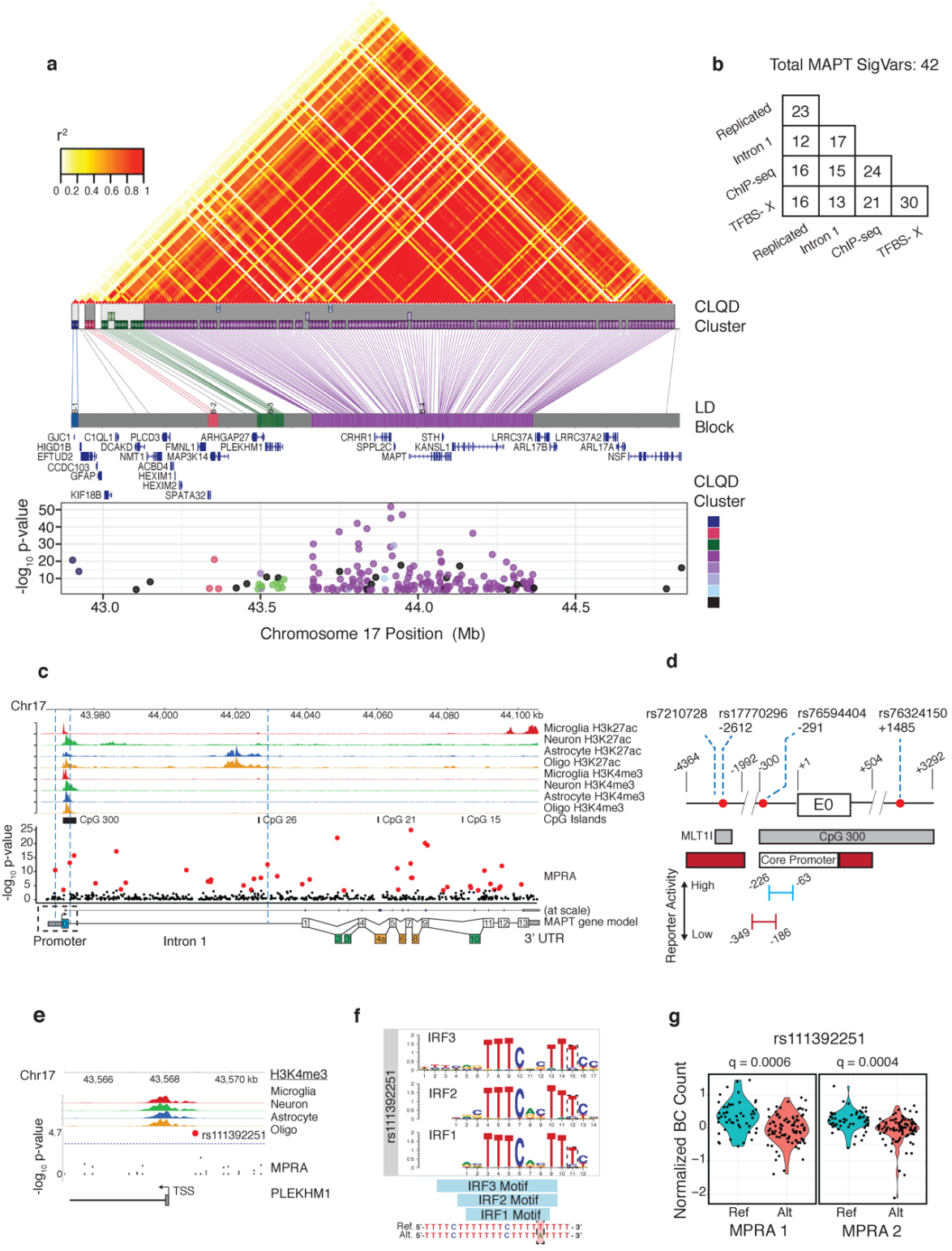


Figure 2-9. Systematic dissection of functional variation at 17q21.31. (**A**) Top: LD plot for common variants in 17q21.31 (1000 Genomes, CEU). MPRA SigVars were clustered by LD (CLQD clustering [41]; Methods). Bottom: SigVars plotted by position across 17q21.31and MPRA significance (-log$_{10}$ p-values; colors = cluster annotations, black variants are unclustered). Most

variants fall within LD-cluster 4 (purple) centering on *MAPT*. (**B**) Annotation breakdown for *MAPT* SigVars. ChIP-seq from [38], TFBS-X = union of predicted TFBS-disruption from two algorithms [37,46]). (**C**) Chromatin annotations and MPRA SigVars across *MAPT*. Genomic tracks (1-8) show H3K27ac (enhancer) and H3K4me3 (promoter) ChIP-seq peaks for microglia (red), neurons (green), astrocytes (blue), and oligodendrocytes (orange) from sorted human brain tissue [38], CpG islands (track 9), and all tested variants plotted by significance ($-\log_{10}$ p-value, track 10). SigVars (FDR q < 0.01, BH method) highlighted in red. Track 11 shows the MAPT gene model. Exons: Untranslated = blue, constitutive = white, alternative neuronal = green, rarely expressed in brain = yellow. (**D**) Functional annotations of the *MAPT* promoter (-4364/+3292, numbering relative to TSS, features from [50]) with SigVars shown (red dots). Red = repressor regions. Bottom: MPRA relative reporter activity for two elements. (**E-G**) highlight rs111392251, a SigVar in the promoter (+/- 3kb) of *PLEKHM1*. (**E**) genomic tracks (1-4) for H3K4me3 peaks (as in **B**). (**F**) The alternate allele of rs111392251 is predicted to significantly disrupt the TFBS for *IRF1-3* TFs. (**G**) violin plots show normalized barcode distributions for each allele (MPRA FDR-adjusted q-values shown).

LD Block	LD Clusters	Number of SigVars	cis Genes
1	4	2	GJC1, HIGD1B, EFTUD2
2	3	3	MAP3K14-AS1, SPATA32, MAP3K14
3	2, 5	12	ARHGAP27, PLEKHM1, LRRC37A4P
4	1,6,7	151	MAPK8IP1P2 ,LINC02210, CRHR1, MAPT-AS1, SPPL2C, MAPT, STH, KANSL1, ARL17B, LRRC37A
NA	NA	20	NA

Table 2-3. SigVars in 17q21.31 grouped by LD block and LD cluster with annotated *cis* Genes (+/- 10 Kb). NA = unclustered variants. SigVars were clustered on the basis of LD using individuals of European ancestry (see Methods), and clusters were assigned into contiguous LD blocks. *Cis* genes are annotated as falling within 10 Kb of any variant assigned to the LD block.

19q.13.32: The *APOE* locus on 19q13.32 harbors the strongest common genetic association with late onset Alzheimer disease (LOAD), tagging the well-characterized *APOE4* risk haplotype [6,19,20]. However, the extensive LD in the region coupled with the strength of the association signal has resulted in identification of hundreds of additional disease-associated variants [18,19]. Recent work involving transethnic scans and haplotype-aware conditional analyses have uncovered evidence for *APOE*- independent risk in the locus, implicating *PVRL2* and *APOC1* [18,53], though others have argued that *APOE* coding variants mediate the entire association

signal in the locus [19,20,54]. I reasoned that identification of functional variants at 19q13.32 may help shed light on this complex regulatory architecture. I tested 640 variants in LD with the 538 genome-wide significant SNPs [6] at this locus, and identified 37 SigVars. Of these, at least 10 were whole blood or brain eQTLs for *PVRL2/NECTIN2* (GTEx)[55]. These loci contained three intronic SigVars (rs34278513, rs412776, rs12972156) that were previously associated with LOAD when analysis was conditioned on *APOE4*-status [18]. My data identifies these variants as causal at this locus, and nominates their target gene as *PVRL2*. I also find that rs141622900, previously associated with cholesterol efflux capacity [56], is a SigVar residing in an active microglial enhancer directly downstream of *APOC1*, providing further support for *APOC1* as an AD risk gene. Finally, I identified an intergenic variant (rs2927437) within a robustly supported multi-tissue enhancer closest to the *BCL3* gene [39].

Discussion

Predicting functionality of noncoding variation is one of the major challenges in modern genetics. In this work, I provide the first systematic characterization of common variants underlying disease risk for two distinct neurodegenerative disorders: Alzheimer's disease and Progressive Supranuclear Palsy. To do so, I designed and implemented two massively parallel reporter assays to screen 5,706 variants encompassing 34 unique loci identified across three genome-wide association studies [6,7,10], and obtained robust activity measures from 94% of tested elements. Most saliently, I identify 320 variants (6% of total) with significant transcriptional skew between alleles at a conservative false discovery rate ($q < 0.01$), thus delineating putative causal variants at 27 of 34 tested genomic loci. Detecting causal variants is critical towards identifying relevant risk genes, or even modeling genetic risk in cellular or *in vivo* systems.

Interestingly, I find that a majority of tested loci contain at least two variants with predicted transcriptional regulatory effects (Figure 2-8c). Previous studies have identified widespread allelic heterogeneity, or the presence of multiple causal variants at a single trait-associated locus. However, the proportions of loci with AH were estimated at 23-35% for two well-powered GWAS [57], less than the 50% identified here. This discrepancy could be due to either MPRA-determined false positives or overly conservative joint-causal variant estimation.

SNP-based heritability is known to be enriched within regulatory regions within disease-relevant tissues [58–60]. In this study, I find that regions of genomic DNA that drive transcriptional activity (*e.g.* "active elements") are significantly enriched within functional regulatory annotations in HEK293T cells known to impact transcription, including open chromatin (DHS) and histone marks delineating active enhancers (H3K27ac; Figure 2-6). This is in agreement with findings from previous MPRAs incorporating alternative library designs and cellular contexts [28,29], and represents an important validation of my screen. Interestingly, I find that approximately ~40% of library elements that overlap DHSs are transcriptionally "active", again in agreement with previous studies that find between 30-46% of surveyed enhancers exhibit functional regulatory activity [61,62]. These results reinforce that while regulatory variation is highly enriched within functional genomic annotations, only a minority of regulatory regions defined by any given annotation are likely to have true transcriptional regulatory effects.

It has been previously shown that trait heritability and GWAS variation enrich within functional annotations for trait-relevant cell or tissue-types [14,58–60]. Indeed, such analyses have been crucial for identifying which cell types mediate common genetic risk, an important step towards interpreting gene mapping studies (Chapter 1). Similar to previous work partitioning heritability in neurologic disorders [63], I demonstrate that MPRA-defined functional regulatory

variants are enriched within active, open chromatin of brain cell types. However, I further show that this remains true of regulatory variants even relative to other GWAS variants within the locus LD block (Figure 2-8e). Moreover, across two separate disorders, I find that a large proportion of functional variants fall within enhancers of distinct cell-types with a demonstrated relationship to disease, and that a large proportion of these enhancers are cell-type specific (Figure 2-8g, 71%). For AD, a plurality of SigVars fall within microglial enhancers, while in PSP a plurality of variants fall within neuronal enhancers. This is in agreement with previous studies partitioning heritability in these disorders [45,64], and suggests that despite a convergence on tau pathology, underlying degenerative mechanisms might be mediated by distinct cellular processes for these two disorders. It should be noted that for PSP, much of this enrichment is due to variants falling within the 17q21.31 region. Unfortunately, there were too few SigVars remaining after excluding variants from the 17q21.31 region to determine whether this neuronal enrichment holds for the other PSP loci.

The enrichment of MPRA SigVars within open chromatin of brain cell-types is also significant because it supports the external validity of this screen. The brain is an organ of extensive cellular heterogeneity, containing divergent cell types from distinct developmental lineages, each playing unique functional roles. I show that there is poor overlap between open chromatin for these cell types, and that HEK293T cells provide an adequate compromise for performing the MPRA within a single cell line. It should be emphasized that this cell line showed the best overlap in active regulatory regions with multiple neuronal and glial cell types that contribute to disease risk, better than either a pure neuronal or glial cell type. Nevertheless, it is reassuring that SigVars identified in an MPRA performed in HEK293T cells ultimately enrich within annotations of the appropriate brain-specific cell types. However, it should be noted that

the specific cellular context and *trans*-acting factors undoubtedly play an important role in the results of functional genomic assays [65]. I addressed this by integrating cell-type specific regulatory annotations to prioritize variants likely to be functional within relevant brain cell types (Supplemental Table 1) to identify high confidence causal variants, as well as performing orthogonal validation using gene editing (Chapter 5). Nevertheless, ideally this screen would be performed in parallel within neurons, glia, and oligodendrocytes, though this is technically challenging (limitations further discussed in Chapter 6).

GWAS loci in regions harboring extensive LD and numerous risk genes are particularly difficult to interpret, and therefore substantially benefit from functional analyses. Here, I used MPRA to identify candidate regulatory variants within two such regions, 17q21.31 and 19q13.32, risk factors for PSP and AD respectively. The 17q21.31 region harbors the tau-encoding *MAPT* gene, which has been genetically linked by familial mutations to tauopathies including PSP and FTLD-tau. Therefore, it is thought that common genetic risk in the 17q21.31 locus might be mediated by haplotypic differences resulting in divergent tau regulation. Indeed, protective H2-haplotype carriers are reported to have decreased brain tau expression and differential splicing of exons 3 and 10 relative to carriers of the H1-risk allele [66]. In agreement with this, I identified numerous variants with significant allelic skew in *MAPT*. Although in contrast to the published work, H2-derived alleles in aggregate increased reporter expression relative to H1 alleles. This may be an artifact of performing the assay in HEK293T cells (Chapter 6); previous studies using reporter assays found *MAPT* promoter variants to have opposite effects depending on technical factors, such as cell type and minimal promoter definition [50].

What is less clear is if genetic risk for tauopathies or PD is mediated by other genes besides *MAPT* within the large 17q21.31 region. Previous work using differential methylation and gene expression analysis have postulated that *LRRC37A*-family or *ARL17A/B* may be candidate risk genes [66,67]. Here, I identify regulatory variants within 29 distinct risk coding genes or lncRNAs at 17q21.31 (Table S2-1). I find PSP-associated regulatory variants within independent LD clusters, representing possible independent causal signal overlapping multiple novel risk genes. One salient example is *PLEKHM1*, a critical regulator of endosome-lysosome fusion and autophagy [52]. I identified rs111392251 in the promoter of *PLEKHM1*, which is in tight LD ($r^2 = 0.93$) with rs11012 - a well-known PD risk variant [68] tagging the H1 haplotype. Similarly, the presence of risk genes independent of APOE within the 19q13.32 AD-risk locus has remained controversial. Here, I identified regulatory variants within 19q13.32 likely impacting expression of *PVRL2/NECTIN2* and *APOC1*. This includes 10 SigVars linked to *PVRL2* on the basis of cell-type specific eQTLs, and a variant in a microglial-specific regulatory region proximal to *APOC1*. Notably, *PVRL2* and *APOC1* have been suggested as *APOE*-independent AD-risk genes [18,69,70] with APOC1 shown to be differentially expressed in microglia derived from AD-brains [71]. Interestingly, I identified two variants in high-confidence enhancers upstream of *BCL3*, an inflammatory TF that is dramatically upregulated in reactive astrocytes downstream of inflammation and in conjunction with AD pathology [72,73]. Further work is needed to validate these regulatory variants within edited isogenic brain relevant cell-lines and to functionally assess the role of these putative risk genes in animal or cellular models of neurodegeneration.

In summary, I utilized massively parallel reporter assays to efficiently characterize variation associated with two neurodegenerative disorders, AD and PSP and

identified 320 variants with significant transcriptional skew between alleles (SigVars). I found a high degree of inter- and intra-library reproducibility, confirming the robustness of these results. Identification of putative causal variants is an important first step towards the mechanistic interpretation of noncoding GWAS loci [74], particularly for complex haplotypes with extended LD such as 17q21.31 and 19q13.32. The MPRA data produced here will also be the basis for subsequent work exploring the genetic architecture and relevant risk genes underlying the neurodegenerative diseases AD and PSP.

Supplement

Chr.	Gene	SigVar Number
chr15	ADAM10	1
chr11	AGBL2	1
chr12	AMHR2	4
chr19	APOC1	1
chr19	APOC1P1	3
chr19	APOC2	4
chr19	APOC4	4
chr19	APOC4-APOC2	4
chr19	APOE	2
chr17	ARHGAP27	7
chr17	ARL17B	15
chr8	ASAP1	1
chr19	BCAM	1
chr19	BCL3	1
chr11	C1QTNF4	1
chr20	CASS4	4
chr19	CBLC	1
chr6	CD2AP	3
chr11	CELF1	1
chr19	CLPTM1	4
chr8	CLU	1
chr1	CR1	2
chr17	CRHR1	11
chr20	CSTF1	2
chr17	CYB561	1

chr17	DCAKD	1
chr18	DSG2	2
chr17	EFTUD2	1
chr7	EPHA1	1
chr7	EPHA1-AS1	3
chr19	EXOC3L2	4
chr11	FAM180B	1
chr14	FERMT2	2
chr11	FNBP4	1
chr17	GJC1	1
chr17	HIGD1B	1
chr6	HLA-DRB1	5
chr6	HLA-DRB4	4
chr6	HLA-DRB6	4
chr2	INPP5D	2
chr16	IQCK	15
chr17	KANSL1	37
chr17	KANSL1-AS1	6
chr11	KBTBD4	1
chr16	KNOP1	5
chr17	LINC02210	3
chr17	LINC02210-CRHR1	45
chr1	LINC02257	2
chr15	LOC101928725	2
chr17	LRRC37A	5
chr17	LRRC37A4P	16
chr17	MAP3K14	3
chr17	MAP3K14-AS1	2
chr17	MAPK8IP1P2	8
chr17	MAPT	45
chr17	MAPT-AS1	21
chr17	MAPT-IT1	8
chr15	MINDY2	2
chr17	MIR4315-1	4
chr17	MIR4315-2	4
chr8	MIR6843	1
chr19	MIR8085	1
chr1	MR1	2
chr11	MS4A4E	1
chr11	MS4A6A	1
chr11	NDUFS3	1
chr19	NECTIN2	16
chr19	NKPD1	1

chr17	NMT1	1
chr17	NSF	2
chr17	PLEKHM1	11
chr19	PPP1R37	2
chr12	PRR13	1
chr8	PTK2B	2
chr6	RUNX2	1
chr14	SLC24A4	1
chr12	SP1	3
chr17	SPATA32	1
chr17	SPPL2C	7
chr17	STH	10
chr1	STX6	1
chr19	TOMM40	3
chr19	TRAPPC6A	1
chr16	VPS35L	3
chr17	WNT3	1

Table S2-1. SigVar annotations by gene. Genes for which there was an MPRA SigVar within +/- 10 kb of the gene body. "SigVar Number" is the number SigVars/gene. Some SigVars are annotated to multiple nearby genes.

Bibliography

1. Prince, M., Guerchet, M. & Prina, M. *The global impact of dementia 2013-2050*. (Alzheimer's Disease International, 2013).
2. Spillantini, M. G. & Goedert, M. Tau pathology and neurodegeneration. *Lancet Neurol.* **12**, 609–622 (2013).
3. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).
4. Forrest, S. L. *et al.* Heritability in frontotemporal tauopathies. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **11**, 115–124 (2019).
5. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088 (2009).

6. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452 (2013).

7. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

8. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

9. Höglinger, G. U. *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699 (2011).

10. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

11. Sanchez-Contreras, M. Y. *et al.* Replication of progressive supranuclear palsy genome-wide association study identifies SLCO1A2 and DUSP10 as new susceptibility loci. *Mol. Neurodegener.* **13**, 37 (2018).

12. Bowles, K. *et al.* 17q21. 31 sub-haplotypes underlying H1-associated risk for Parkinson's disease and progressive supranuclear palsy converge on altered glial regulation. (2019).

13. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).

14. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

15. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).

16. Pittman, A. M. *et al.* The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum. Mol. Genet.* **13**, 1267–1274 (2004).

17. Goedert, M. & Jakes, R. Mutations causing neurodegenerative tauopathies. *Biochim. Biophys. Acta BBA-Mol. Basis Dis.* **1739**, 240–250 (2005).

18. Zhou, X. *et al.* Non-coding variability at the APOE locus contributes to the Alzheimer's risk. *Nat. Commun.* **10**, 1–16 (2019).

19. Jun, G. *et al.* Comprehensive search for Alzheimer disease susceptibility loci in the APOE region. *Arch. Neurol.* **69**, 1270–1279 (2012).

20. Naj, A. C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* **43**, 436–441 (2011).

21. 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).

22. Liu, L. *et al.* Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 1–11 (2019).

23. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).

24. 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).

25. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

26. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).

27. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

28. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

29. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

30. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 1–14 (2017).

31. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).

32. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).

33. Sørensen, A. T. *et al.* A robust activity marking system for exploring active neuronal ensembles. *Elife* **5**, e13918 (2016).

34. Bushnell, B., Rood, J. & Singer, E. BBMerge–accurate paired shotgun read merging via overlap. *PloS One* **12**, (2017).

35. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).

36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

37. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS–a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).

38. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).

39. Wang, J. *et al.* HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).

40. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015 (2015).

41. Kim, S. A. *et al.* gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics* **35**, 4419–4421 (2019).

42. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. Massively parallel reporter assays in cultured mammalian cells. *JoVE J. Vis. Exp.* e51719 (2014).

43. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 1–15 (2019).

44. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. (2012).

45. Swarup, V. *et al.* Identification of conserved proteomic networks in neurodegenerative dementia. *Cell Rep.* **31**, 107807 (2020).

46. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).

47. Caillet-Boudin, M.-L., Buée, L., Sergeant, N. & Lefebvre, B. Regulation of human MAPT gene expression. *Mol. Neurodegener.* **10**, 1–14 (2015).

48. Andreadis, A., Wagner, B. K., Broderick, J. A. & Kosik, K. S. A τ promoter region without neuronal specificity. *J. Neurochem.* **66**, 2257–2263 (1996).

49. Sadot, E., Heicklen-Klein, A., Barg, J., Lazarovici, P. & Ginzburg, I. *Identification of a tau promoter region mediating tissue-specific-regulated expression in PC12 cells*. (Academic Press, 1996).

50. Maloney, B. & Lahiri, D. K. Structural and functional characterization of H2 haplotype MAPT promoter: unique neurospecific domains and a hypoxia-inducible element would enhance rationally targeted tauopathy research for Alzheimer's disease. *Gene* **501**, 63–78 (2012).

51. Heicklen-Klein, A. & Ginzburg, I. Tau promoter confers neuronal specificity and binds Sp1 and AP-2. *J. Neurochem.* **75**, 1408–1418 (2000).

52. McEwan, D. G. *et al.* PLEKHM1 regulates autophagosome-lysosome fusion through HOPS complex and LC3/GABARAP proteins. *Mol. Cell* **57**, 39–54 (2015).

53. Kulminski, A. M., Philipp, I., Loika, Y., He, L. & Culminskaya, I. Haplotype architecture of the Alzheimer's risk in the APOE region via co-skewness. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **12**, e12129 (2020).

54. Jun, G. R. *et al.* Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimers Dement.* **13**, 727–738 (2017).

55. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

56. Low-Kam, C. *et al.* Variants at the APOE/C1/C2/C4 Locus Modulate Cholesterol Efflux Capacity Independently of High-Density Lipoprotein Cholesterol. *J. Am. Heart Assoc.* **7**, e009545 (2018).

57. Hormozdiari, F. *et al.* Widespread allelic heterogeneity in complex traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).

58. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

59. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

60. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

61. 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).

62. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).

63. Consortium, B. Analysis of shared heritability in common disorders of the brain. *Sci. N. Y. NY* **360**, (2018).

64. Swarup, V. *et al.* Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nat. Med.* **25**, 152–164 (2019).

65. Shigaki, D. *et al.* Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.* **40**, 1280–1291 (2019).

66. Li, Y. *et al.* An epigenetic signature in peripheral blood associated with the haplotype on 17q21. 31, a risk factor for neurodegenerative tauopathy. *PLoS Genet* **10**, e1004211 (2014).

67. Allen, M. *et al.* Gene expression, methylation and neuropathology correlations at progressive supranuclear palsy risk loci. *Acta Neuropathol. (Berl.)* **132**, 197–211 (2016).

68. Edwards, T. L. *et al.* Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.* **74**, 97–109 (2010).

69. Ki, C.-S., Na, D. L., Kim, D. K., Kim, H. J. & Kim, J.-W. Genetic association of an apolipoprotein CI (APOC1) gene polymorphism with late-onset Alzheimer's disease. *Neurosci. Lett.* **319**, 75–78 (2002).

70. Zhou, Q. *et al.* Association between APOC1 polymorphism and Alzheimer's disease: a case-control study and meta-analysis. *PloS One* **9**, e87017 (2014).

71. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

72. Zamanian, J. L. *et al.* Genomic analysis of reactive astrogliosis. *J. Neurosci.* **32**, 6391–6410 (2012).

73. Srinivasan, K. *et al.* Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat. Commun.* **7**, 1–16 (2016).

74. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

CHAPTER 3

Performance of algorithms for computational variant prediction

Abstract

The functional annotation of noncoding genetic variation within trait-associated genomic loci is a major problem in modern genetics, motivating the development of both fine-mapping approaches and computational algorithms for variant prioritization. Machine learning methods leveraging features such as functional genomic annotations and conservation scores to predict variant pathogenicity have been recently popularized, though the relationship between predicted pathogenicity and empirically determined regulatory function remains unclear. I therefore compared the predicted scores from four such algorithms, CADD, CATO, GWAVA, and LINSIGHT with empirical effect sizes for common noncoding variants derived from two separate MPRA datasets. Generally, computational algorithms failed to predict MPRA-determined variant functionality, though the CATO algorithm showed modest correlation with MPRA effect sizes. Moreover, computational algorithms were highly discordant with each other. Finally, I found that computational algorithms struggled to functionally discriminate between variants that were closely spaced together suggesting an overreliance on spatially broad genomic annotations during algorithm training. These results indicate that empirical assays provide orthogonal measures of variant function, which may aid in the future validation and improvement of predictive algorithms.

Introduction

The functional annotation of noncoding variation is an outstanding problem in human genetics, an issue that has become particularly salient due to the widespread adoption and utilization of the Genome Wide Association Study (GWAS). This poses a significant barrier impeding biological interpretation of GWAS loci, motivating the development of a number of

statistical and computational approaches for functional variant prioritization. These methods can be divided into two main flavors: fine-mapping [1] and computational prioritization algorithms [2].

Fine-mapping encompasses a set of statistical methods that use association statistics to identify causal variants. These approaches include penalized regression or probabilistic Bayesian models and often incorporate simplifying underlying assumptions such as the number of causal variants, prior causal distribution, and local LD structure derived from population panels. Some approaches also integrate genomic annotations or gene expression to weight prior distributions or further narrow credible sets [1]. Bayesian methods, including PAINTOR [3], eCAVIAR [4], and SuSiE [5] are currently favored, having been shown to have improved performance through detailed simulation studies. Another fine-mapping strategy is the trans-ethnic scan [6,7]. This approach is founded upon the observation that GWAS results are reproducible across populations, with the expectation that underlying causal variants are shared. Therefore, cross-population meta-analysis would greatly reduce average locus LD and subsequently improve detection power for causal variants. However, fine-mapping approaches are subject to a number of limitations. Notably, they require sufficiently large GWAS sample sizes to have adequate power. Furthermore, these methods perform poorly in regions of strong LD; in the extreme case of complete LD it is impossible to disambiguate variants. Finally, these approaches are incompatible with prioritization of rare or structural variants. For a detailed review, readers are directed to the work of Schaid and colleagues [1].

In contrast to fine-mapping approaches, computational prioritization algorithms are agnostic to association statistics and LD structure, instead considering SNPs on an individual basis [2]. Earlier methods, including VEP [8], RegulomeDB [9], and FunciSNP [10], primarily use functional annotations such as chromatin accessibility, ChIP-seq, eQTLs, and TFBS annotations

to prioritize variants (Chapter 1). Other methods also incorporate evolutionary conservation as a predictive feature, under the assumption that constrained nucleotides have a higher probability of functional relevance [11,12]. This assumption is tenuous in rapidly evolving noncoding regulatory domains such as human accelerated regions or enhancers for genes under positive selection [13]. Finally, machine learning approaches incorporate functional annotations, conservation, nucleotide composition, and other features as predictors in models trained on curated or simulated sets of variants [14–18]. For a comprehensive review of machine learning methods for GWAS prioritization, readers are directed to Nicholls and colleagues [19] .

The choice of variation used to train machine learning models is critical to their theoretical performance, which is problematic given that there is no gold standard dataset of functional noncoding variants [2]. This issue similarly extends to method validation and benchmarking. Often, classification performance is based on discrimination of pathogenic variants from curated databases (such as ClinVar [20] or HGMD [21]) or disease- associated variation from GWAS. This approach can introduce bias through implicit distributional assumptions made on control variants or through clinical ascertainment, skewing performance measurement. Some methods explicitly train on databases of pathogenic variants (e.g. GWAVA [17] and FATHMM-MKL [18]) introducing the possibility of overfitting. MPRA data, which directly measures the transcriptional-regulatory effects of polymorphisms, represents a truly orthogonal data-type which can be used as an alternative benchmarking strategy for these computational approaches. Indeed, there are a few examples in the literature correlating scores from predictive algorithms with MPRA data and results have been disparate. For example, Nishizaki and Boyle find some correlation (Pearson's r^2 = 0.17-0.3) between four methods (DeepSEA, FATHMM-MKL, RegulomeDB, CADD) and variants identified by MPRA as disrupting function of liver

enhancers[2]. Similarly, Ulirsch and colleagues tested variants derived from GWAS of red blood cell traits and found strong correlation with predicted effects from DeepSEA [22]. By contrast, in an analysis of 19 different methods by Kircher and colleagues, computational predictions explained a very small proportion of the variance in MPRA data (average Pearson's r = 0.03) [23].

Previously, I used MPRA to characterize 5,706 noncoding variants derived from GWAS for two neurodegenerative disorders, Alzheimer's disease and Progressive Supranuclear Palsy (Chapter 2). Therefore, I reasoned that I could use these data to provide further insight into the relationship between computational prediction algorithms and MPRA experimental outcomes, adding to this growing literature. Here, I compare pre-computed scores derived from four commonly used algorithms, CADD [24], GWAVA [17], LINSIGHT [25], and CATO [26] with my MPRA results, identifying generally low concordance between different prioritization methods.

Materials and Methods

MPRA datasets

For this analysis I utilized two distinct MPRA datasets. The first was my MPRA dataset (see Chapter 2 for full description of the Methods). Summary statistics (p-values and $\log_2$ Fold Changes) were combined across both MPRA stages, and a total of 5,340 unique variants were considered. The second dataset was a previously published analysis of lymphoblastoid eQTLs performed in K562 cells by Tewhey and colleagues [27]. In this study, some variants were tested in multiple orientations or with multiple genetic backgrounds. I filtered this dataset to get one effect-size per individual variant, by keeping the most significant statistical comparison per variant (max negative log p-value).

Comparison with computational predication algorithms:

I scored all tested variants from both datasets using the LINSIGHT, CADD, CATO, and GWAVA algorithms [17,24–26], using each algorithm's precomputed scores. When an algorithm provided multiple scores per variant (particularly LINSIGHT, which provides scores in small genomic windows that sometimes overlapped), scores were averaged. First, I compared scores from the top vs bottom 5th percentile of variants as ranked by descending -log MPRA p-value using a Mann-Whitney-U test. The "top" variants were those with the most significant allelic skew, while "bottom" were least significant. I then correlated (Spearman's rho) MPRA absolute value effect-sizes (i.e. absolute value $\log_2$ Fold Change) with computational predicted scores. This was done using all the variants, only variants with significant allelic skew (SigVars), and SigVars + 100 bottom ranked variants.

I then labeled MPRA SigVars at FDR-adjusted $q < 0.01$ thresholds and calculated Area Under the Receiver Operating Curve for each algorithm using the ModelMetrics (1.2.2.2) package in R (v. 4.0.0). I then performed binary classification on all scored variants by positively labeling MPRA SigVars ($q < 0.01$) and an equivalent number of the top scoring variants from each algorithm before calculating pairwise Cohen's Kappa using the ModelMetrics package.

Finally, I tested whether there was an increased similarity between scores from variants that were nearby in genomic space compared with distantly spaced variants as follows: Variants were ordered by chromosome and position. Then, for each variant [V_i], I determined the absolute value difference in score between [V_i] and [V_{i+1}]. A similar analysis was done using MPRA effect sizes instead of scores. These absolute score differences were binned as "near" if the genomic distance between [V_i] and [V_{i+1}] was less than 100 base pairs, or "far" if otherwise. I then compared "near" vs "far" score differences using a two-tailed Mann-Whitney-U test.

Statistical analysis and data visualization

All statistical analysis was performed using the stats package in R. All tests were two-sided. Data were visualized using ggplot2.

Results

I compared MPRA-determined effect sizes and allelic skew significance levels - derived from this study and the previously published MPRA dataset [27] - with regulatory scores from four widely used algorithms: CADD, CATO, GWAVA, and LINSIGHT [17,24–26]. For both MPRA datasets, representing vastly different disorders and tissue types, I observed that these computational methods indeed were able to capture enrichment of regulatory signal. I compared the top vs bottom 5th percentile of variants as ranked by allelic skew p-values, and found a significant (two-sided Mann-Whitney-U test) increase in average algorithm prediction score for the top ranked MPRA variants across all algorithms in at least one study (Figure 3-1).

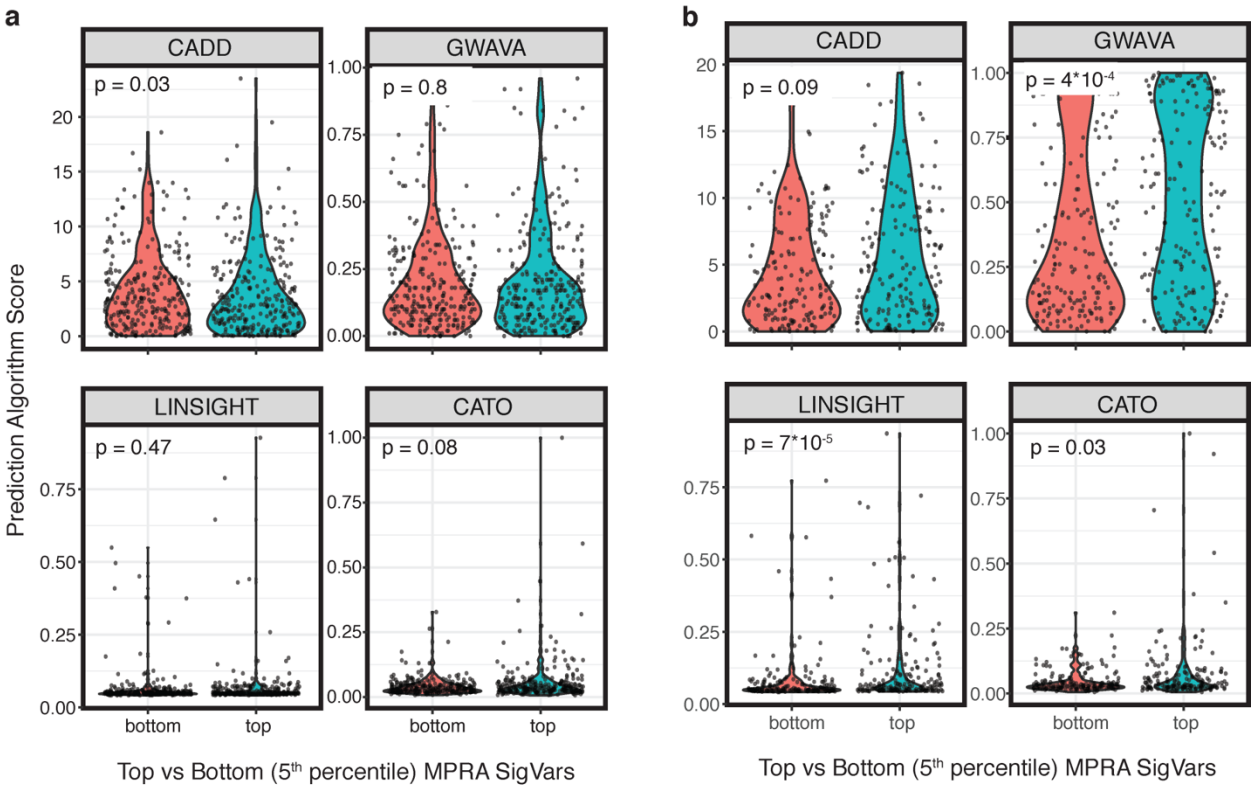


Figure 3-1. Top-ranked MPRA variants enrich for increased functional prediction scores. All MPRA-tested variants were scored using four variant-effect prediction algorithms; LINSIGHT, CADD, GWAVA, and CATO. Violin plots show algorithm prediction scores for the top vs. bottom 5th percentile of ranked variants (rank determined by MPRA allelic skew p-values) in the current study (**A**) and replication dataset [27] (**B**). p-values from two-sided Mann-Whitney-U test.

However, if I applied a statistical threshold that would identify variants with functional effects (allelic skew FDR-corrected thresholds < 0.01; Methods), overall regulatory predictions were highly discordant and not strongly predictive (max AUC = 0.55, 0.56; Figure 3-2a-b). MPRA effect sizes also failed to correlate with algorithm predictions, with the exception of CATO (Pearson's r = 0.14 – 0.19, both p < 0.001; Figure 3-2c-d). Previous studies have described correlations between MPRA effect sizes and algorithm prediction scores. In these studies, only variants with significant allelic skew [22] or a collection composed of mostly significant variants and a handful of negative controls [2] were compared with algorithm predictions. I therefore considered that my approach might be overly conservative by including a

large proportion of non-significant variants with expected null effect sizes. However, the correlations between CADD, GWAVA, and LINSIGHT scores and MPRA effect sizes for: 1) only SigVars or, 2) a collection of SigVars and 100 non-significant variants, remained nonsignificant (all $p > 0.05$) for both my dataset and the replication dataset. Correlations with CATO improved somewhat from baseline when only considering SigVars ($r = 0.23$-0.29).

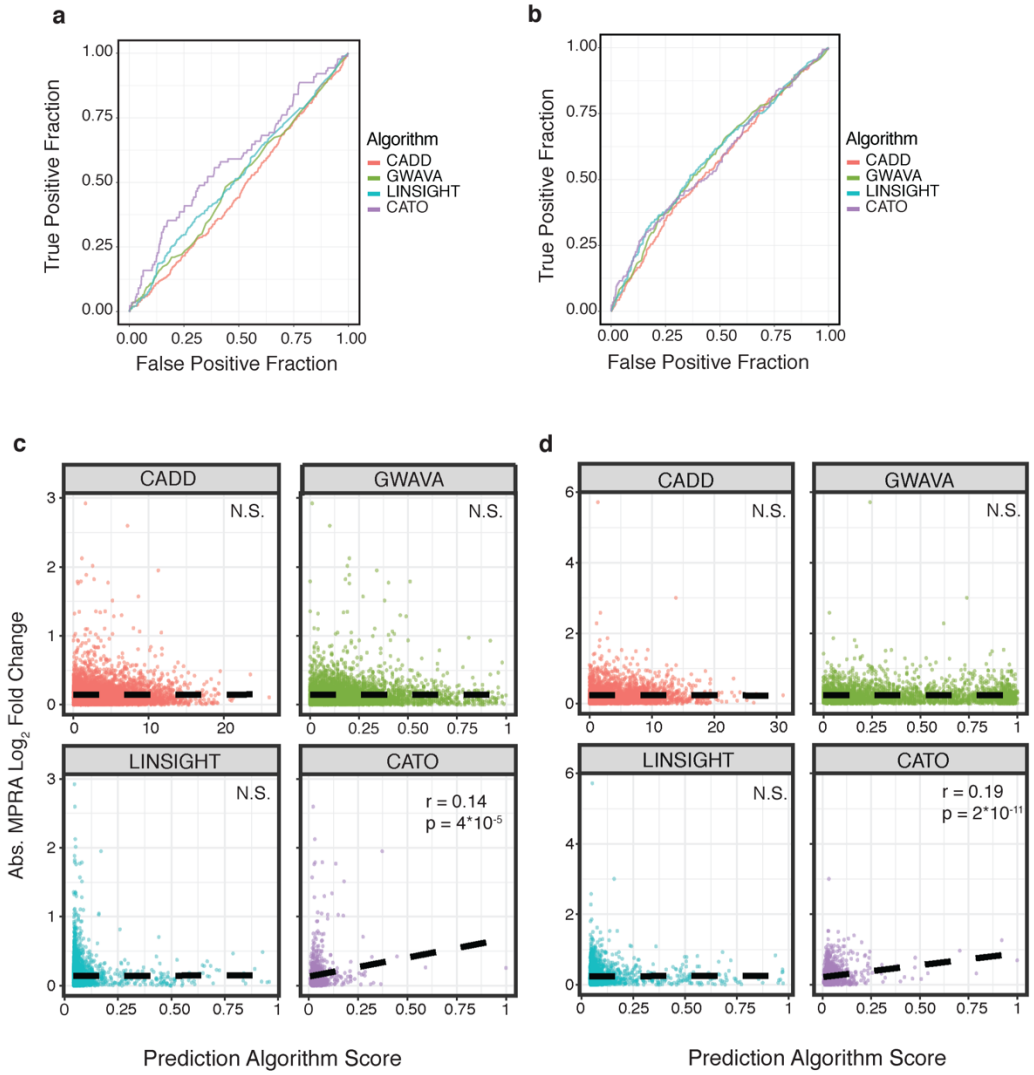


Figure 3-2. Computational prediction algorithms poorly predict MPRA empirical effect sizes. **A**) ROC curves highlight the poor predictive performance (max AUC = 0.55) of four algorithms used to score variant functionality (LINSIGHT, CADD, GWAVA, and CATO) benchmarked against MPRA SigVars (defined at $q < 0.01$, BH method). (**B**) same as (**A**) except using SigVars from a published MPRA dataset [27] (max AUC = 0.56). When comparing all tested variants, prediction algorithm scores correlated poorly with MPRA-determined allelic skew effect sizes in

both my dataset (**C**) and the replication dataset (**D**) with the exception of the CATO algorithm. Black dashed line = OLS regression line of best fit. Pearson's r, all p > 0.1 (N.S.), except for CATO algorithm.

I also calculated pairwise overlap metrics and found that MPRA SigVar experimental outcomes had little overlap with computational predictions (mean Cohen's Kappa = 0; Figure 3-3; Methods). Interestingly, prediction algorithms had little concordance among themselves; overall mean Kappa between all predicative algorithms was 0.11 for variants in our dataset and 0.14 for variants from the other published MPRA [27].

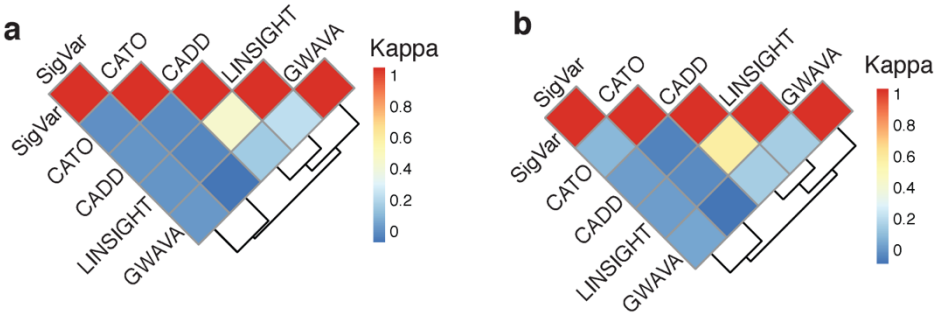


Figure 3-3. Pairwise Cohen's Kappa for MPRA SigVars and variant prediction algorithms. Top-ranked variants scored by each algorithm (Methods) for this study (**A**) and the published study [27] (**B**).

I next considered whether computational prediction algorithms could functionally discriminate between closely spaced variants. To do so, I computed the difference in predicted score between every variant and its nearest neighbor, and then binned these score differences into "near" (variants < 100 base pairs apart) and "far" (> 100 bp apart) categories. I then compared the average difference in scores between these two categories for all algorithms using a Mann-Whitney-U test (Methods). I found that for all algorithms, variants that were closely spaced together had more similar predicted pathogenicity scores (i.e. smaller average differences) than variants that were far apart. These results suggest that algorithm predictions are biased by physical proximity, translating into a difficulty in discriminating closely spaced

regulatory variants. By contrast, the difference between MPRA effect sizes remained consistent irrespective of variant distance (Figure 3-4).

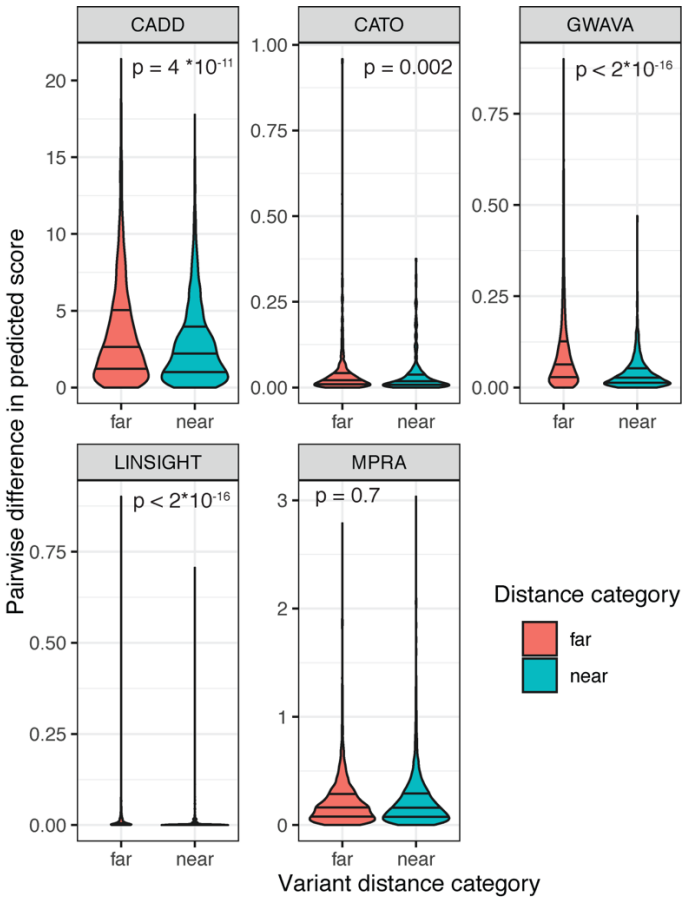


Figure 3-4. Effect of variant proximity on algorithm functional predictions. The difference in algorithm prediction scores (or MPRA effect sizes) was computed pairwise for all variants and their nearest neighbor. These scores were binned into "near" or "far" categories if the genomic distance between variant pairs was less than or greater than 100 base pairs respectively. Violin plots display the score difference distribution by category for each algorithm, with quantiles demarcated. Two-sided Mann-Whitney-U test p-values are shown.

Discussion

The use of computational methods for *a priori* identification of functional classes of variants is an ongoing area of active research and several predictive methods have been developed. I show here that four commonly used machine learning methods for variant functional prediction, CADD [24], GWAVA [17], LINSIGHT [25], and CATO [26], poorly recapitulate

my MPRA data and subsequent variant annotations, as well as previously published MPRA data in a different cell type from Tewhey *et al* [27]. Of note, predictions from these four methods were also highly discordant with each other (Figure 3-3), an observation that has been noted repeatedly in the literature [28].

Machine learning methods can be broken down into three classes [25]: 1) Methods, including FATHMM-MKL [18] and GWAVA [17], that are trained to classify curated sets of known deleterious mutations using genomic annotations. 2) Methods such as CADD [24], DANN [15], LINSIGHT [25], and FitCons2 [11] that use genomic features and inferred patterns of selection to predict the effects of mutations on evolutionary fitness. 3) Methods such as CATO [26], DeepSEA [14], and deltaSVM [16] that dispense of predicting variant pathogenicity, and instead identify sequence based predictors of cell-type specific annotations (such as DNA-accessibility or ChIP-seq). All of these can be further contrasted with unsupervised learning methods such as EIGEN which infer variant classes based on spectral decomposition of a genomic annotation matrix [29]. Considering these diverse objective functions (i.e. pathogenicity, selection, molecular phenotypes), it is perhaps unsurprising that method predictions are highly discordant, especially with MPRA data which is measuring a separate phenotype of transcriptional regulation. It is therefore likely that these different approaches have advantages or weaknesses depending on the specific prediction task. For example, approaches that predict variant effects on evolutionary fitness (e.g. LINSIGHT) are not limited by the same ascertainment bias as methods trained on curated sets of variants, but may be less accurate in regions of strong positive selection or on traits that are not under selective constraint (e.g. post-reproductive phenotypes) [25]. Methods that learn sequence effects on molecular phenotypes depend on excellent cell-type specific genomic annotations, and make strong assumptions about the relevance of these phenotypes on the trait of

interest. This corresponds with the observation that different methods have highly variable performance at predicting MPRA variant activity across different genes and classes of noncoding genomic regions [23]. Indeed, recent omnibus methods attempt to boost performance by integrating scores across many algorithms, though benchmarking against MPRA data has remained relatively poor [30].

One weakness consistent across many types of methods is a reliance on functional genomic annotations for model training. This manifests as poor nucleotide-level predictive resolution [28], and can be obscured by deceptive performance benchmarking. Method performance is often evaluated based on discrimination of specific sets of variants. For example, a task might be to identify GWAS variants against an allele-frequency matched set of controls from flanking genomic regions. As I have described previously (Chapter 1), GWAS variation is highly enriched within functional genomic elements and key genic annotations [31,32]. It is therefore unsurprising that predictive methods trained on these annotations achieve high AUROC (which can be a misleading measure [33]) given the relative genomic sparsity of these features. Likewise, methods will train on conservation scores (such as SIFT [34], PhyloP [35], GERP++[36]) which are expectedly higher in genic regions, and then "demonstrate" predictive performance by showing enrichment of predicted pathogenic variants within conserved regions with obvious functional consequences (e.g., promoters, UTRs, splice sites, etc.)[25]. Therefore, while these methods can demonstrate enrichment with functional genomic elements, they struggle to disambiguate functional variants *within* these elements, which are often hundreds of base pairs wide. It has been shown previously that six predictive methods perform poorly at prioritizing closely spaced variants, or different allelic substitutions at sites [28]. Similarly, I show here that variants that are close together are more likely to have similar pathogenicity scores.

This is not the case for MPRA, which directly measures variant effects, and may be another explanation for the poor correlation between these methods and MPRA.

Methods that articulate sequence or motif-level predictors (CATO [26], deepSEA [14], deltaSVM [16], gkm-SVM [37]) might therefore be expected to have the best single-nucleotide resolution and highest fidelity to MPRA data, which indeed has been reported [2,22,23,30]. Similarly, in this study, the only algorithm that correlated with MPRA outcomes was CATO, which explicitly models TFBS-occupancy. Indeed, TFBS annotations are an important predictor of MPRA performance (Chapter 4). Another advantage of these approaches is that they incorporate cell-type specific information, as the cellular environment is a critical component of reporter assay performance [38]. In a prior study using gkm-SVM, prediction of MPRA variant function in mouse retina was highly dependent on model training using retina-specific annotations [39]. Interestingly, performance was further improved by incorporating MPRA training data, which suggests that such data types could be used to further improve algorithm predictive performance in the future.

Bibliography

1. 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).
2. Nishizaki, S. S. & Boyle, A. P. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends Genet.* **33**, 34–45 (2017).
3. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).

4. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

5. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

6. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).

7. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

8. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).

9. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

10. Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A. & Noushmehr, H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* **40**, e139–e139 (2012).

11. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).

12. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).

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

14. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).

15. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).

16. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955 (2015).

17. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of non-coding sequence variants. *Nat. Methods* **11**, 294–296 (2014).

18. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).

19. Nicholls, H. L. *et al.* Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front. Genet.* **11**, 350 (2020).

20. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

21. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).

22. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

23. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 1–15 (2019).

24. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

25. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).

26. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393 (2015).

27. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

28. Liu, L. *et al.* Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 1–11 (2019).

29. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214 (2016).

30. Zhang, S. *et al.* regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. *Nucleic Acids Res.* **47**, e134–e134 (2019).

31. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

32. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

33. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).

34. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073 (2009).

35. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).

36. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).

37. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**, e1003711 (2014).

38. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts. *Biol. Psychiatry* (2020).

39. Beer, M. A. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* **38**, 1251–1258 (2017).

CHAPTER 4

Functional variation disrupts disease-specific transcriptional networks

Abstract

Transcription factor binding sites are key genomic features mediating transcriptional regulation. Despite technical advances in the genome-wide prediction of transcription factor binding sites (TFBS), the gene-regulatory effects and downstream phenotypic relevance of genetic variation overlapping these motifs remains unclear. Massively Parallel Reporter Assays (MPRA) can efficiently survey the effects of polymorphisms on gene regulation. I previously used MPRA to screen genetic variation derived from GWAS for two neurodegenerative disorders, Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP). Here, I utilize two algorithms to predict the functional effects of this same variation on transcription factor binding. I find that predicted TFBS-alteration scores significantly correlate with empirical variant effect sizes and that disruption of *ETS* and *SP* -family TFBSs most strongly predicts MPRA outcomes. Moreover, I find that MPRA-defined functional variation preferentially disrupts binding sites for particular transcription factors that differ drastically by disease. Binding sites for these disease-specific transcription factors enrich within regulatory regions with differential cell-type specific activity in PSP and AD, which specifically implicates dysregulation of a neuronal *SP1*-driven transcriptional network in PSP pathogenesis. These analyses support a novel mechanism underlying noncoding genetic risk, whereby common genetic variation distributed across the genome functions in aggregate to drive disease risk via dysregulation of specific transcriptional programs.

Introduction

The binding of transcription factors to their cognate genomic binding sites is a key step during transcriptional initiation [1]. Therefore, characterizing the spatiotemporal binding dynamics and sequence specificities of the more than 1,600 identified human transcription factors is critical towards understanding gene regulation and the impact of genetic variation [2]. Modern high-throughput approaches including ChIP-seq, HT-SELEX, and DAP-seq [3] efficiently survey physical protein-DNA interactions at the genome scale, and these data can subsequently be used to infer TF sequence specificities [4]. These binding specificities are often represented in a position weight matrix (PWM) summarizing nucleotide probabilities at each position in the binding motif, and curated collections of TF PWMs are available in online repositories including TRANSFAC [5], JASPAR [4], and HOCOMOCO [6]. A number of algorithms have been developed that incorporate PWMs to scan DNA sequences and identify binding motifs in a statistically rigorous manner, which can be used to efficiently predict TFBSs genome-wide [7,8]. However, it is estimated that less than 1% of annotated TFBSs are actually bound by TFs *in vivo* [1], indicating that genomic annotation of core motifs is necessary but not sufficient to predict functionality. Indeed, TF-DNA binding is complex and is moderated by distal sequence context, binding of co-factors, genomic accessibility, TF expression, and DNA methylation [9].

The modification of TFBSs within gene regulatory elements is an important mechanism whereby genetic variation may contribute to disease risk. Across traits, GWAS variation has been found to be enriched within gene regulatory elements and TFBSs [10,11], and there are numerous examples linking specific motif-disrupting variants with gene expression and phenotypic alterations [9]. This has motivated the development of a number of *in silico* variant functional profiling tools that predict the effects of genetic variation on TF binding [12,13].

However, the proportion of trait variance attributable to direct disruption of TFBSs - in contrast to other underlying mechanisms - has remained unclear. This is due to the abovementioned difficulties in identifying TFBSs with true functional effects in relevant cell-types, assigning causal variants within a given GWAS locus, and determining whether predicted motif alterations translate to true binding disruption *in vivo* [9,14]. Indeed, it has been observed in TF ChIP-seq data that only a minority of allele-specific binding events were attributable to direct sequence alterations of the core TF-binding motif [15], pointing to the importance of broader sequence context and cooperative binding dynamics. Experimental approaches such as Massively Parallel Reporter Assays that directly assess variant effects in a cellular context have begun to answer these questions.

Previous studies utilizing MPRA have highlighted the significance of TFBSs on reporter function. For example, in one analysis of 56 regulatory features, TF-motif density was the single most predictive feature of MPRA activity [16]. However, the relationship between TFBS disruption and assay performance remains less well characterized. I previously used MPRA to screen 5,706 polymorphisms derived from GWAS for two neurodegenerative disorders, and found a significant enrichment of TF binding motifs within active regulatory elements (Chapter 2). In this work, I find a complex but significant relationship between predicted TFBS-disruption and MPRA-determined allelic effects for these disease associated variants. I subsequently find that functional variants are enriched within the binding sites of specific transcription factors that form cell-type and disease-specific regulatory networks.

Materials and Methods

MPRA datasets

For this analysis I utilized two distinct MPRA datasets. The first was my MPRA dataset (see Chapter 2 for full description of the Methods). Summary statistics (p-values and $\log_2$ Fold Change) were combined across both MPRA stages, and a total of 5,340 unique variants were considered. The second dataset was a previously published analysis of lymphoblastoid eQTLs performed in K562 cells by Tewhey and colleagues [17]. In this study, some variants were tested in multiple orientations or with multiple genetic backgrounds. I filtered this dataset to get one effect-size per individual variant, by keeping the most significant statistical comparison per variant (max negative log p-value).

TFBS analysis

Correlation with MPRA effect sizes: TFBS disruption was scored for all variants from my MPRA dataset using the SNPS2TFBS webtool [12]. An enrichment odds ratio for predicted TFBS-disruption between variants with and without significant allelic skew (FDR-adjusted $q < 0.01$) was calculated using Fisher's exact test. For SigVars predicted to disrupt TFBSs, I correlated allelic skew effect sizes from my MPRA with predicted TFBS disruption scores. Sites with a TFBS score of 0 (denoting poorly defined scores) were discarded and the score with the max absolute value was chosen for sites with multiple predicted disruptions. I performed a similar analysis for SNPs from another large MPRA dataset [17]. As this dataset characterized variants embedded in genomic context from both positive and negative strands, I used the max observed effect size.

TFBS-disruption enrichment: I then partitioned SigVars into those that were derived from AD or PSP GWAS (111 and 209 variants respectively; 17q21.31 variants were considered PSP), and re-ran them through the SNPS2TFBS algorithm which outputs enrichment odds ratios for each TF based on their background binding site probability in the genome. This output was filtered for TFBSs with at least two predicted disruptions and enrichment p-values were then FDR-adjusted (BH method). Only TFBSs significantly ($q < 0.05$ or $q < 0.1$) enriched for disruption by SigVars were plotted using ggplot2, with disruption counts, $\log_2$ enrichment odds ratios, and unadjusted -$\log_{10}$ p-values shown (Figure 4-3). As a negative control I also tested random samples of 111 or 209 variants to calculate the expected null distribution.

Correlation with TF abundance: I also tested whether HEK293T cell TF abundance (gene expression) correlates with MPRA allelic skew. To do so, I scored all variants tested in MPRA 1 using the SNPS2TFBS software. For each variant predicted to disrupt a TFBS, I found the corresponding TF's normalized gene expression in HEK293T cells using RNA-seq data from the human protein atlas [18] (accession in Appendix A – Supplemental Materials and Methods). I then found the Spearman's correlation between TF expression and the MPRA absolute value $\log_2$ fold change.

Protein-Protein Interaction network construction: For the 6 TFs enriched for significant binding site disruption in PSP (at $q < 0.05$ threshold) or the 14 TFs at $q < 0.1$, I created a protein-protein interaction network using the STRING (v11) webtool and standard parameters (Figure 4-3) [19]. I then computed empirical SP1-connectivity p-values for this network using resampling as follows: I determined protein-protein interactions for all TFs annotated by the SNPS2TFBS tool

(165 TFs total) using STRING (https://string-db.org). I then found the number of edges between SP1 and either 5 or 13 additional randomly sampled TFs, repeating this procedure 10,000 times to create a distribution, which was compared to the true number of PSP-network edges (4 or 8 respectively) to generate a SP1-connectivity p-value. I then identified all sites among the 209 PSP-SigVars predicted to disrupt binding of these 6 TFs (union of SNPS2TFBS and motifbreakR annotations; Supplemental Table 1) and found all genes within +/- 10 kb of these sites, which I called "target genes" (Figure 4-4). I then annotated the network composed of these TFs and their target genes using single cell RNA-seq data from human M1 cortex (© 2010 Allen Institute for Brain Science. Allen Human Brain Atlas. Available from: https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x) [20]. I annotated each network gene as belonging to the corresponding cell type with the highest trimmed mean gene expression value, and created a boxplot displaying the number of annotated genes per cell type.

TFBS-disruption predicts MPRA activity: I also tested whether TFBS disruption was predictive of MPRA allelic skew. For this analysis, I considered variants derived from the external MPRA dataset [17]. Only variants with computed allelic skew values were considered. I used the maximum $-\log_{10}$ p-value for variants tested in multiple configurations and also filtered out variants missing valid rsIDs. I then performed FDR adjustment (BH method) on the MPRA-determined allelic skew p-values, and labeled SigVars at $q < 0.01$ thresholds. All variants were scored for TFBS disruption using the motifbreakR package (2.2.0) utilizing the HOCOMOCO v10 TF binding model [6,13] ("strong" effect, and binding threshold of $p < 1\times10^{-4}$) .Variants that could not be identified in dbSNP v144 were not scored and were subsequently discarded. I then computed the positive predictive value between predicted TFBS disruption and SigVar labels. I

also downloaded RNA-seq data for K562 cells from the human protein atlas [18], and identified the top 200 TFs (by normalized counts) expressed in these cells. I then filtered for "strong" predicted TFBS disruption only for motifs corresponding to these top expressed TFs, and then re-computed PPD.

Finally, I tested whether disruption of particular TFBSs predicted MPRA allelic skew (Figure 4-2). I again only considered TFs expressed in K562 cells. All variants were previously scored for TFBS-disruption (described above) and were grouped according by specific TF annotation. For each collection of variants predicted to disrupt a particular TFBS, I found the proportion of these variants that were also annotated as SigVars (SigVars defined at $q < 0.01$). I then assessed the statistical significance of these proportions for each TF by performing a one-tailed binomial test against the background SigVar probability in the overall dataset (prob = 0.115), leaving a p-value which was FDR adjusted.

Statistical analysis and visualization

Statistical analysis was performed using the stats package in R (version 4.0.0). All data were plotted using the ggplot2 package.

Results

MPRA SigVars enrich for transcription factor binding site disruption

I previously found that active MPRA elements were enriched for TFBSs (Chapter 2). I therefore hypothesized that functional variants, defined as variants with significant transcriptional skew between alleles (*i.e.* SigVars), would be enriched for variants that disrupt TF-binding as a class. Therefore, I ran all my previously screened variants through the

SNPS2TFBS algorithm [12], which predicts TFBS disruption, finding that SigVars were enriched for variants that disrupt TFBSs (OR = 1.4, p = 0.003) compared to variants without significant allelic skew. I observed a similar magnitude of enrichment of TFBS-disrupting variants amongst SigVars from a previously published dataset from Tewhey *et al.* [17] (OR = 1.9, $p = 8.7 \times 10^{-8}$). These findings were reproduced in both my dataset (OR = 1.77, $p = 9.8 \times 10^{-7}$) and the previously published dataset (OR = 1.70, $p = 1.3 \times 10^{-6}$) using an alternative TFBS-scoring method (motifbreakR; Methods). Furthermore, the magnitude and directionality of predicted TFBS disruption correlated with MPRA effect sizes in both my dataset (Spearman's rho = 0.44, $p = 2.3 \times 10^{-4}$) and the published dataset (rho = 0.52, $p < 8.7 \times 10^{-9}$; Figure 4-1).

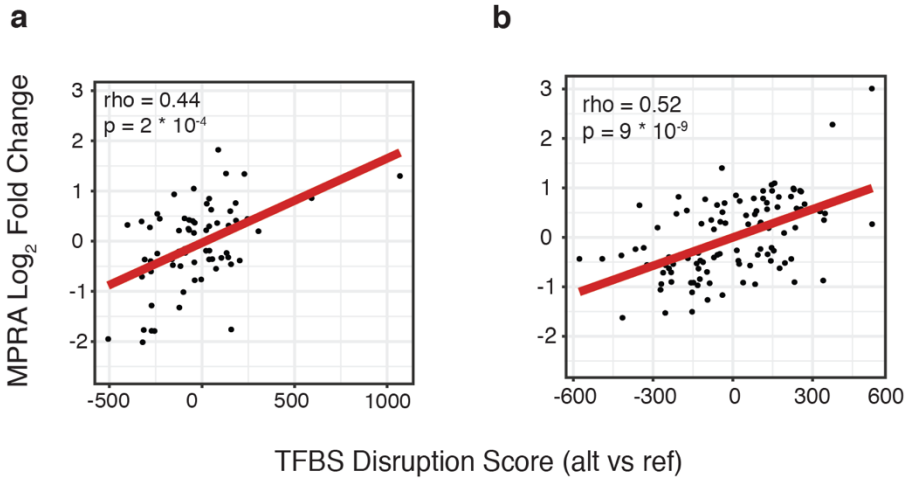
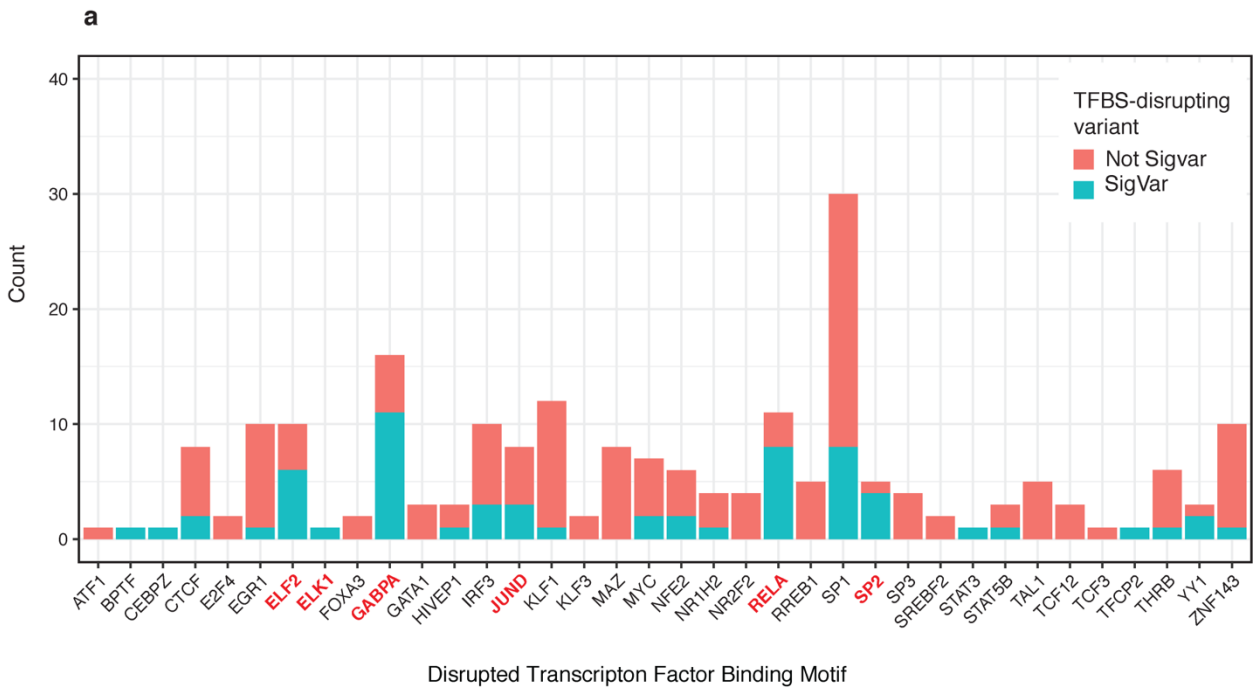


Figure 4-1. MPRA variant effect sizes correlate with predicted TFBS disruption. MPRA $\log_2$ fold changes significantly correlate with TFBS-disruption scores from the SNPS2TFBS algorithm in my dataset (**A**) and the replication dataset (**B**). OLS line of best fit shown in red, Spearman's rho also shown.

Given this enrichment, I asked whether TFBS disruption alone predicts MPRA allelic skew. To do so, I scored all variants from the previously published dataset [17] (assay performed in K562 cells) for TFBS disruption (Methods) and found that only a proportion of TFBS-disrupting variants were also MPRA SigVars (Positive Predictive Value = 0.14). Predictive performance somewhat improved after filtering to only consider strongly disrupted binding sites for the top

200 K562-expressed TFs (PPD = 0.22; Methods). This predictive performance is on par with the top-performing algorithm, CATO [21] (Chapter 3). These findings demonstrate the importance of accounting for cell-type specific *trans*-factors such as transcription factor expression profiles for variant functional predictions. Nevertheless, although the overlap between predicted TFBS disruption and MPRA SigVar annotation was significantly higher than chance, TFBS-disruption only explains a subset of MPRA-defined functional variants. Interestingly, TFBS disruption was more predictive for some TFs over others. To test this, I grouped together all variants predicted to disrupt binding of a particular TF, and then found the proportion of variants in each group that were also MPRA SigVars (Methods). I found variants that disrupted certain TFBSs were highly enriched for MPRA SigVars. For example, disruption of *ELK1*, *ELF2*, and *GABPA* TFBSs (all ETS-family TFs), as well as SP-family TFBSs, were highly predictive of allelic skew captured by MPRA (Figure 4-2), while disruption of other TF families was not.



b

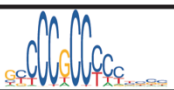
TF	Family	Motif	p-values	q-values
GABPA	Ets-related		3.5e-12	1.7e-10
RELA	NF-kB -related		1.4e-5	0.0003
ELF2	Ets-related		6.5e-5	0.0001
JUND	JUN		0.0005	0.006
ELK1	Ets-related		0.002	0.02
SP2	3-finger Kruppel-related SP - subfamily		0.002	0.02
SP1	3-finger Kruppel-related SP - subfamily		0.05	0.3

Figure 4-2. TFBS disruption is predictive of MPRA allelic skew for specific classes of TFs, including the ETS-related TF family. (**A**) Variants with significant allelic skew (SigVars) were defined at an FDR-adjusted threshold of $q < 0.01$ (BH method) for variants tested in a previously published large MPRA dataset. These variants were also assessed for predicted TFBS disruption (motifbreakR) filtering for motifs corresponding to the top 200 most highly expressed TFs in K562 cells. The histogram counts the number of TFBS-disrupting variants annotated for each TF, colored by whether variants are also MPRA SigVars (blue) or not (red). TFs for which disrupted binding significantly predicts MPRA allelic skew are highlighted in red, and also

described further in (**B**). Determination of a significantly increased proportion of SigVars amongst variants predicted to disrupt specific TFBSs was assessed using a one-sided binomial test against the background SigVar probability, with p-values and FDR-adjusted q-values shown (Methods).

Enrichment of functional risk variants within disease-specific transcriptional networks

I next assessed whether SigVars were enriched within binding sites for specific transcription factors. Importantly, I determined that TFBSs disrupted by risk variants differed by disease. In AD, *NR4A2* ($\log_2$ OR = 4.9, p = 0.002), *NR5A2* ($\log_2$ OR = 3.6, p = 0.01), *ATOH1* ($\log_2$ OR = 3.8, p = 0.009), *SP2* ($\log_2$ OR = 2.3, p = 0.008) , and *SMAD*-family *(SMAD2,3,4* heterotrimer, $\log_2$ OR = 3.7, p = 0.0002) binding sites were enriched for disrupting risk variants (all enrichments at FDR-adjusted $q < 0.05$). Interestingly, PSP showed a different pattern of TFBS enrichment, in which five of the six TFs predicted to be enriched for binding site disruption physically interact with the transcription factor SP1 (including SP1 itself; Figure 4-3a). Importantly, while these TFs are highly expressed in HEK293T cells, I did not find a general relationship between relative TF abundance and MPRA allelic skew ($p > 0.05$; Spearman's correlation; Methods), indicating that expression levels in HEK293T cells do not explain these results. Additionally, as a negative control I ran equivalently sized random samples of variants through the SNPS2TFBS algorithm (Methods), only identifying one TF enrichment per sample (TFs: *BACH1* and *STAT6*), demonstrating that the above results are not generic to any random collection of GWAS variants. Finally, I also assessed enrichments at a lower statistical threshold (TFBS enrichment $q < 0.1$), finding a number of ETS-family TFs and SP1 interactors to be enriched within functional PSP risk variants (Figure 4-3b).

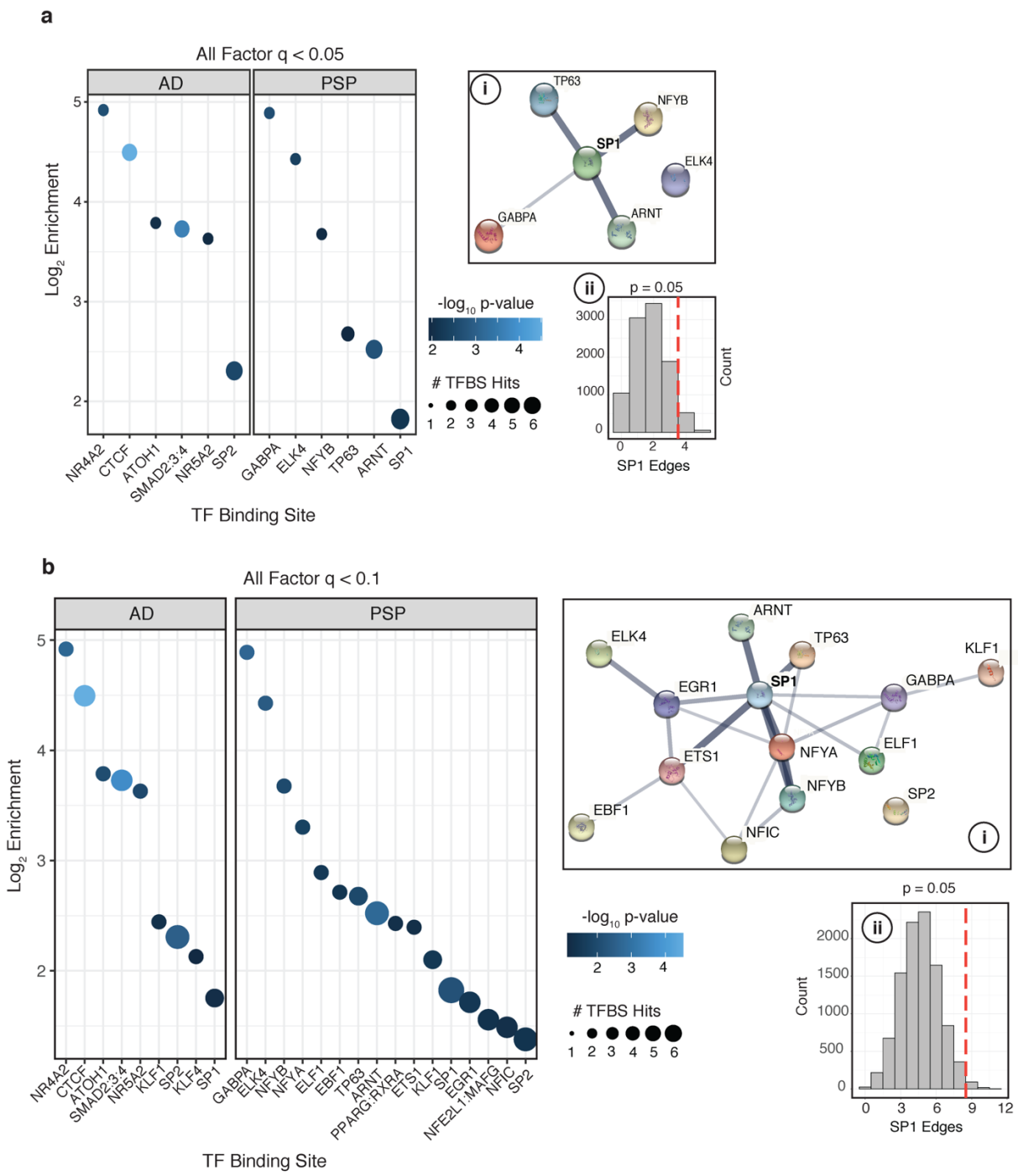


Figure 4-3. MPRA SigVars are enriched for TFBS-disruption in a disease-specific manner. (**A**) Shows the $\log_2$ enrichments for significantly disrupted TFBSs (FDR q < 0.05, BH method), with SigVars from AD and PSP analyzed separately (color = $-\log_{10}$ enrichment p-values, size = # of disrupted TFBSs; SNPS2TFBS [12]). Inset: i) protein-protein interaction network from the STRING [19] database for significantly disrupted PSP-TFs. Line thickness = strength of evidence. ii) Empirical distribution for expected PPI network SP1 – node connectivity generated by resampling (Methods). Red lines = observed PSP-TF network edges, with permutation p-value shown. (**B**) is the same as (**A**), except enrichment threshold was defined at FDR q < 0.1.

Further analysis indicated that the PSP-enriched TFs form a significant protein-protein interaction network with SP1 (permutation p = 0.05; Figure 4-3 insets; Methods), consistent with SP1's multimerization capabilities and its activity as a core component of a broad array of gene regulatory complexes that regulate tissue specific gene expression [22]. Interestingly, 11.1% of annotated PSP SigVars are predicted to participate specifically within this network (20% when including factors from the expanded network). Of note, the coding region for *SP1* itself falls within a suggestive PSP risk locus approaching genome-wide significance (locus combined p-value = 4.1×10^{-7}) [23]. These data, including identification of five functional regulatory variants at the *SP1* locus, provide strong evidence for disruption of an SP1-based signaling network in PSP pathophysiology.

To explore this further, I next generated a two-layer directed network composed of these six significant PSP TFs ($q < 0.05$) and their likely targets, defined as genes within the *cis*-regulatory window of TFBS-disrupting SigVars (Figure 4-4a; Methods). Using available single-cell human brain gene expression data [20], I annotated all members of this network for their highest expressed cell type (Methods), and found that all but one of these genes were expressed most strongly in neurons (Figure 4-4b), suggesting that this disrupted PSP-associated signaling network functions primarily within neurons, consistent with the observation of overall neuronal enrichment of heritable PSP risk variants from GWAS [24].

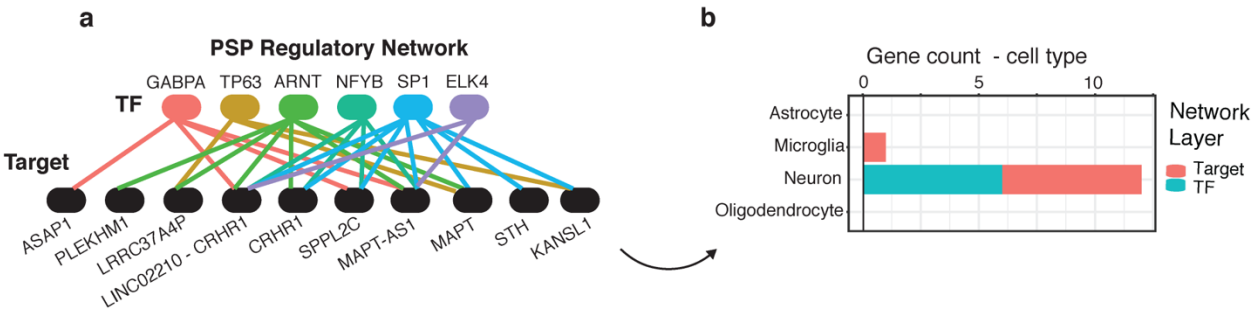


Figure 4-4. A disrupted PSP transcriptional network is most highly expressed in neurons. (**A**) Network of TFs whose binding is significantly disrupted in PSP and their target genes (genes

within 10 kb of disrupted TFBSs). (**B**) Bar plot counts the top-expressing brain cell-type for each gene in (**A**) colored by network layer (cyan = TF, red = target gene).

Discussion

Transcription factor binding sites are critical genomic features mediating transcriptional regulation [1]. Curated databases of TF binding specificities coupled with modern statistical approaches have enabled efficient scanning and prediction of TFBSs genome wide [4,8]. However, the proportion and circumstances in which predicted binding sites might be functionally relevant, as well as the effects of genetic variation overlapping these sites remains uncertain. In this work, I used data from two separate Massively Parallel Reporter Assays to clarify the relationship between predicted TFBS disruption and empirically determined measures of variant effects on transcriptional regulation. First, I confirmed an enrichment of predicted TFBS-disruption amongst variants with significant transcriptional skew between alleles (SigVars), and found strong correlation between MPRA determined effect sizes and predicted TFBS alterations. This confirms previous reports in the literature that TFBS disruption is a strong predictor of MPRA activity [16,17]. Interestingly, I found that certain TF families, including the ETS and SP family TFs are more predictive of MPRA performance than other classes. This also replicates previous reports that identify ETS and SP/KLF family, as well as AP-family TFBSs as being highly predictive of MPRA experimental outcomes in multiple cell types [16,25,26]. Why single nucleotide disruption of core binding motifs predictably impacts transcription for some factors and not others is an unresolved question, but may have to do with baseline binding stringency and specificity, or evolutionary constraint. Whether these findings remain consistent across different experimental designs and cell types also remains to be explored.

I previously generated MPRA data characterizing common variation associated with two neurodegenerative disorders. Here, I find that binding sites for specific TFs are enriched for disrupting functional variation, enabling identification of specific transcriptional networks driving key aspects of disease risk (Figure 4-3). For AD, this involves TFs including *NR4A2* and *SMAD*-family TFs, which have been described previously as acting within multiple cell-types in the brain and periphery to impact risk for AD [27,28]. For PSP, my analysis identified an enrichment of TFBS-disrupting functional variation in a significant protein interaction network with the transcription factor SP1. These TFs, as well as most of their predicted regulatory target genes are most highly expressed in neurons, consistent with cell-type specific aggregation of genetic risk in PSP. Moreover, I identified five SigVars in the PSP genome-wide suggestive locus harboring *SP1* [23], as well as four SP1 binding site-disrupting functional variants within the *MAPT* gene (Chapter 2; Supplemental Table 1). These convergent findings provide strong evidence for disruption of SP1 signaling in neurons as a critical risk factor for PSP. SP1 is known to regulate a broad array of cellular processes including chromatin remodeling, apoptosis, immune regulation, and response to oxidative stress in neurons [29]. While SP1 network dysregulation has been identified in AD brain [30–32], genes within this network do not harbor an over-representation of genetic risk; thus, this network is likely to play a reactive or secondary role. Overall, these data are consistent with PSP risk primarily impacting neurons, and astrocytes and oligodendrocytes secondarily, and AD risk in microglia and astrocytes, as has been reported [24,33].

My observations also suggest a refined model for understanding common genetic risk. Signal from intergenic GWAS loci are typically interpreted as deriving from causal regulatory variants that influence downstream expression of specific cognate risk genes, and thus are explainable through colocalization with eQTLs. My results, implicating a TF network

converging on SP1 in PSP, are consistent with a model whereby common genetic variants function in aggregate across multiple TFBSs to disrupt key cell-type specific transcriptional programs. I speculate that this genetic mechanism may particularly manifest itself only upon induction of the relevant transcriptional network, which may occur within a disease context. For example, transcriptomic and proteomic studies have previously identified induction of an SP1 transcriptional network in tauopathies [31,32], which here I show may interact with common variation at relevant binding sites distributed across the genome to determine risk. It has been previously observed that GWAS loci and eQTLs exhibit limited genetic sharing, leading to a "missing" mechanism of action explaining a large proportion of noncoding loci [34]. My data show that polymorphisms altering binding of critical, disease-relevant transcriptional networks offer an additional explanation. That these transcriptional networks regulate a large number of cell-type enriched genes, provides a mechanism whereby genetic risk is expressed, not by impacting a few core genes [35], but via polygenic cell type-specific regulatory effects on networks of genes [36].

Bibliography

1. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
2. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
3. Bartlett, A. *et al.* Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* **12**, 1659 (2017).
4. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

5. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).

6. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

7. Jayaram, N., Usvyat, D. & Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* **17**, 1–12 (2016).

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

9. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).

10. Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).

11. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

12. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS–a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).

13. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).

14. Moyerbrailean, G. A. *et al.* Which genetics variants in DNase-Seq footprints are more likely to alter binding? *PLoS Genet.* **12**, e1005875 (2016).

15. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22**, 860–869 (2012).

16. Kreimer, A., Yan, Z., Ahituv, N. & Yosef, N. Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. *Hum. Mutat.* **40**, 1299–1313 (2019).

17. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

18. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).

19. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

20. Bakken, T. E. *et al.* Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *bioRxiv* (2020).

21. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393 (2015).

22. Bouwman, P. & Philipsen, S. Regulation of the activity of Sp1-related transcription factors. *Mol. Cell. Endocrinol.* **195**, 27–38 (2002).

23. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

24. Swarup, V. *et al.* Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nat. Med.* **25**, 152–164 (2019).

25. 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).

26. Movva, R. *et al.* Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PloS One* **14**, e0218073 (2019).

27. Moon, M. *et al.* Nurr1 (NR4A2) regulates Alzheimer's disease-related pathogenesis and cognitive function in the 5XFAD mouse model. *Aging Cell* **18**, e12866 (2019).

28. Town, T. *et al.* Blocking TGF-β–Smad2/3 innate immune signaling mitigates Alzheimer-like pathology. *Nat. Med.* **14**, 681–687 (2008).

29. Tan, N. Y. & Khachigian, L. M. Sp1 phosphorylation and its regulation of gene transcription. *Mol. Cell. Biol.* **29**, 2483–2488 (2009).

30. Citron, B. A., Dennis, J. S., Zeitlin, R. S. & Echeverria, V. Transcription factor Sp1 dysregulation in Alzheimer's disease. *J. Neurosci. Res.* **86**, 2499–2504 (2008).

31. Canchi, S. *et al.* Integrating gene and protein expression reveals perturbed functional networks in Alzheimer's disease. *Cell Rep.* **28**, 1103–1116 (2019).

32. Andreev, V. P. *et al.* Label-free quantitative LC–MS proteomics of Alzheimer's disease and normally aged human brains. *J. Proteome Res.* **11**, 3053–3067 (2012).

33. Swarup, V. *et al.* Identification of conserved proteomic networks in neurodegenerative dementia. *Cell Rep.* **31**, 107807 (2020).

34. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

35. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

36. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).

CHAPTER 5

Genome editing validates putative regulatory variants within AD GWAS loci

Abstract

Massively Parallel Reporter Assays are a versatile high-throughput approach enabling the functional characterization of noncoding genetic elements. MPRAs have recently been utilized to successfully screen trait-associated loci derived from genome-wide association studies to identify underlying regulatory variants. However, MPRAs are synthetic assays that test variants removed from their native chromatin context and therefore may be susceptible to false positive results. I previously used MPRA to screen variation derived from GWAS for two neurodegenerative disorders. Here, I use CRISPR/Cas9 technology to validate six putative regulatory variants implicated in that screen by performing genome editing within brain-relevant cell lines. In doing so, I confirm rs13025717 and rs636317 as regulatory variants for *BIN1* and *MS4A6A* respectively, validating two well-known Alzheimer's disease (AD) risk loci. Furthermore, I find that rs6064392 likely increases *CASS4* gene expression within microglia to increase AD risk. Finally, I identify pleiotropic regulatory function for rs9271171, a variant that lies within the complex *HLA-DR* region. Interestingly, I find that this variant modifies expression of the distal gene complement 4 in both monocytic cell lines and IPSC-derived astrocytes, implicating *C4* as a novel genetic risk factor for AD.

Introduction

Over the last few decades, the widespread deployment of genotyping and next generation sequencing technologies have enabled the identification of millions of noncoding genetic variants associated with numerous traits, including neurodegenerative and neuropsychiatric disorders [1]. However, the mechanistic and functional characterization of this variation has remained limited. Transcriptomic, epigenomic, and proteomic profiling of patient-derived tissue

has enabled detailed functional mapping of the noncoding genome forming the basis for subsequent variant prioritization [2] (Chapter 1). However, these studies are correlational in nature, and are also reflective of a single post-mortem snapshot rather than a complex disease trajectory. By contrast, genome editing technologies enable the direct manipulation and functional characterization of noncoding genetic elements. This effort has been revolutionized by the recent identification of an efficient, flexible, and programmable nuclease known as the CRISPR/*Cas* system [3].

CRISPR technology is based on an adaptive antiviral system first identified in bacteria, composed of genetic elements called CRISPRs (clustered regularly interspaced short palindromic repeats) and *Cas* nucleases. CRISPRs encode short single-stranded guide RNAs which interact with Cas proteins to target complementary DNA sequences for cleavage and degradation [4]. The CRISPR system can be harnessed to flexibly target experimental sequences of interest through the rational design of sgRNAs, and the technology was first adapted for application in mammalian systems less than a decade ago [5,6]. At its most rudimentary, CRISPR enables the precise generation of double-stranded breaks at target sites, subsequently eliciting repair using non-homologous end-joining. This pathway is error-prone, often resulting in small deletions and frameshift mutations useful for gene knock-out studies [3]. However, researchers quickly realized that the alternate homologous recombination pathway could be harnessed to introduce precise mutations of interest by providing an exogenous DNA repair template [7]. Later technological iterations fused nuclease-deactivated Cas with the effector domains of other proteins to enable specific editing *sans* template (base editing [8] or prime editing [9]), as well as experimental applications beyond genome editing that include functional manipulation of gene expression (e.g. CRISPRi [10] and CRISPRa [11]), the epigenome [12,13], or chromatin topology [14]. Excitingly,

these assays can be scaled by generating pooled libraries of many gRNAs and combining them with single-cell phenotyping to create high-throughput CRISPR screens [15]. These screens can be used to survey the noncoding regulatory landscape, as highlighted by a recent study that paired CRISPR inactivation with scRNA-seq to characterize close to 6,000 putative enhancers in K562 cells [16]. Notably, CRISPR screens have also been implemented within IPSC-derived neurons [17].

Massively Parallel Reporter Assays are an alternate approach for the high-throughput experimental characterization of noncoding genetic features [18]. However, MPRAs survey genetic elements removed from their natural genomic and chromatin environment, and therefore may not directly recapitulate features in their native context. Previous studies utilizing MPRA to survey natural human variation therefore verified key findings using genome editing within relevant cell-types [19–21]. I previously used MPRA to screen 5,706 common variants associated with Alzheimer's disease and Progressive Supranuclear Palsy (Chapter 2), identifying 320 variants with significant transcriptional skew between alleles (SigVars). Here, I use CRISPR/Cas9 genome editing within brain-relevant cell lines to validate four of six tested SigVars as likely regulatory variants.

Methods

MPRA data used and visualized here was previously generated and described (Chapter 2).

Cell Culture

I obtained HEK293T (CRL-3216), THP-1 (TIB-202), and SH-SY5Y (CRL-2266) cell lines from ATCC. HEK293T cells were cultured in DMEM containing GlutaMAX (Thermo Fisher Scientific, 10566016) supplemented with 10% FBS and 1% Sodium Pyruvate (11360070).

THP-1 cells were cultured in RPMI-1640 medium (ATCC 30-2001) supplemented with 10% heat-inactivated FCS and 10 mM HEPES buffer. THP-1 differentiation was performed through addition of 20 ng/mL of PMA (Millipore Sigma, P1585) to the culture media for 48 hours. SH-SY5Y cells were cultured in 1:1 EMEM/Ham's F12 (ATCC 30-2003 / Thermo Fisher Scientific, 31765035) supplemented with 15% HI-FCS, 1% Sodium Pyruvate, and 1% MEM-NEAA (Thermo Fisher Scientific, 11140050). For differentiation, SH-SY5Y cells were gently passaged and plated on Poly-D Lysine coated plates. After 24 hours culture media was replaced with a differentiation media composed of Neurobasal A (Thermo Fisher Scientific, 10888022), GlutaMAX, B27 Supplement (17504044), and 10 uM Retinoic Acid (Millipore Sigma, R2625). Cells were differentiated for 5 days with half-media replacement every 48 hours.

Induced Pluripotent Stem Cells (IPSCs) used in this study were previously documented [22] and kindly provided by Dr. Li Gan in accordance with the UCLA TDG guidelines. IPSCs were differentiated into mature astrocytes for 120 days as previously described [23], and were maintained post-differentiation in DMEM supplemented with 1% Sodium Pyruvate, 10% HI-FCS, and 1x N2 supplement (Thermo Fisher Scientific, 17502048) until use.

CRISPR experiments

I excised enhancers containing rs636317, rs13025717, rs6064392, rs9271171, rs1532277, and rs7920721 as follows: Pairs of guide RNAs targeting upstream (5') and downstream (3') flanking sequences were designed and cloned into LentiCRISPRv2-GFP (Addgene #82416) and LentiCRISPRv2-mCherry (#99154) respectively using the BsmbI restriction site (gRNA sequences in Appendix A – Supplemental Materials and Methods). Lentiviral particles were produced in HEK293T cells by triple transfection as previously described

(https://www.addgene.org/protocols/lentivirus-production), concentrated using Lenti-X (Takara Bio, 631232), and resuspended in DPBS. Guide pairs were then screened for cutting efficiency resulting in two pairs targeting rs13025717 (gRNA #s: 11 + 12, 11 + 14), one pair targeting rs636317 (15 + 16), one pair targeting rs6064392 (25 + 26), three pairs targeting rs9271171 (3 + 4, 5 + 4, 5 + 6), three pairs targeting rs1532277 (7 + 8, 7 +10, 9 + 10), and one pair targeting rs7920721 (21 + 20). Guide pairs or scramble gRNA control lentiviruses (MOI ~ 0.5) were then used to infect 80% confluent 6-well plates of SH-SY5Y cells or t25 flasks of THP-1 cells. Culture media was replaced 16 hours later and cells were expanded for 5 days post-infection. Cells were sorted at the UCLA BSCRC flow cytometry core to isolate ~500,000 GFP+/mCherry+ cells per replicate, and were subsequently differentiated. Ultimately I excised a 240 or 374 bp region containing rs13025717 in SH-SY5Y cells, a 430 bp region containing rs636317 in THP-1 cells, a 382 bp region containing rs6064392 in THP-1 cells, a 1065 or 682 bp region containing rs9271171 in THP-1 cells (as well as a 1005 bp region in IPSC-derived astrocytes), a 393, 259, or 200 bp region containing rs1532277 in SH-SY5Y cells, and a 473 bp region surrounding rs7920721 in THP-1 cells.

For each replicate I collected total RNA and genomic DNA using the AllPrep DNA/RNA Mini Kit (Qiagen, 80204). CRISPR-mediated removal of the target enhancer was assessed by amplifying the target region of gDNA by PCR (genomic PCR primers: Appendix A – Supplemental Materials and Methods) and verifying strong representation of the truncated allele via gel electrophoresis. cDNA was reversed transcribed using SuperScript IV, Oligo$(dT)_{20}$ primer, and 300ng of total RNA. We performed qPCR using the KAPA SYBR FAST Kit (Roche, KK4600), 500nM qPCR primers (qPCR primers: Appendix A – Supplemental Materials

and Methods) and a Roche LightCycler 480. Relative transcript abundance was quantified using the $2^{-\Delta\Delta CT}$ method [24] normalized to the geometric mean of ACTB and GAPDH reference genes.

Statistical analysis and data visualization

Statistical analysis was performed using the stats package in R (version 4.0.0). Group comparisons were performed using a two-sided Student's t-test. Genomic tracks and chromatin annotations were visualized using the pyGenomeTracks python module [25]. All other data were plotted using the ggplot2 package.

Results

Validation of select MPRA SigVars

Beginning with the 320 significant putative regulatory variants identified in Chapter 2 at a conservative false discovery rate (FDR-adjusted $q < 0.01$), I used brain functional regulatory annotations to identify 55 high-confidence likely causal variants (Chapter 2; Methods). Here, I selected six of these variants for additional validation, including three variants within well-characterized AD loci near *BIN1, CLU,* and *MS4A6*, and three variants within less well-described loci near *HLA-DRB1/5*, *CASS4*, and *ECDHC3*. I used CRISPR-Cas9 genome editing to assay regulatory regions in their native genomic context by excising the enhancer elements containing these variants, then assaying downstream effects on gene expression using quantitative PCR (Methods). I identified the target genes for a given variant either when the variant was close to the gene within its cis-regulatory region, or those genes linked by chromatin interaction data based on Hi-C from a relevant cell-type.

BIN1 gene expression is regulated by rs13025717

I first assessed rs13025717, which is a highly significant MPRA SigVar (MPRA 1 q = 2.6×10^{-38}). This variant resides about ~20 kb from the transcriptional start site (TSS) of the established AD risk gene *BIN1*. It is also predicted to strongly disrupt binding of the transcription factor KLF4, and overlaps functional microglial and monocyte annotations including DNAse Hypersensitivity Sites (DHS) and H3K27ac, H3K4me1, and H3K4me3 ChIP-seq peaks (Figure 5-1a-c). I used two pairs of gRNAs to excise a 240 or 374 bp segment containing rs13025717 in the SH-SY5Y cell line, and found a significant reduction in *BIN1* gene expression (Figure 5-1d). Of note, I also tried performing this experiment in differentiated THP-1 cells, but *BIN1* was not expressed in this cell type.

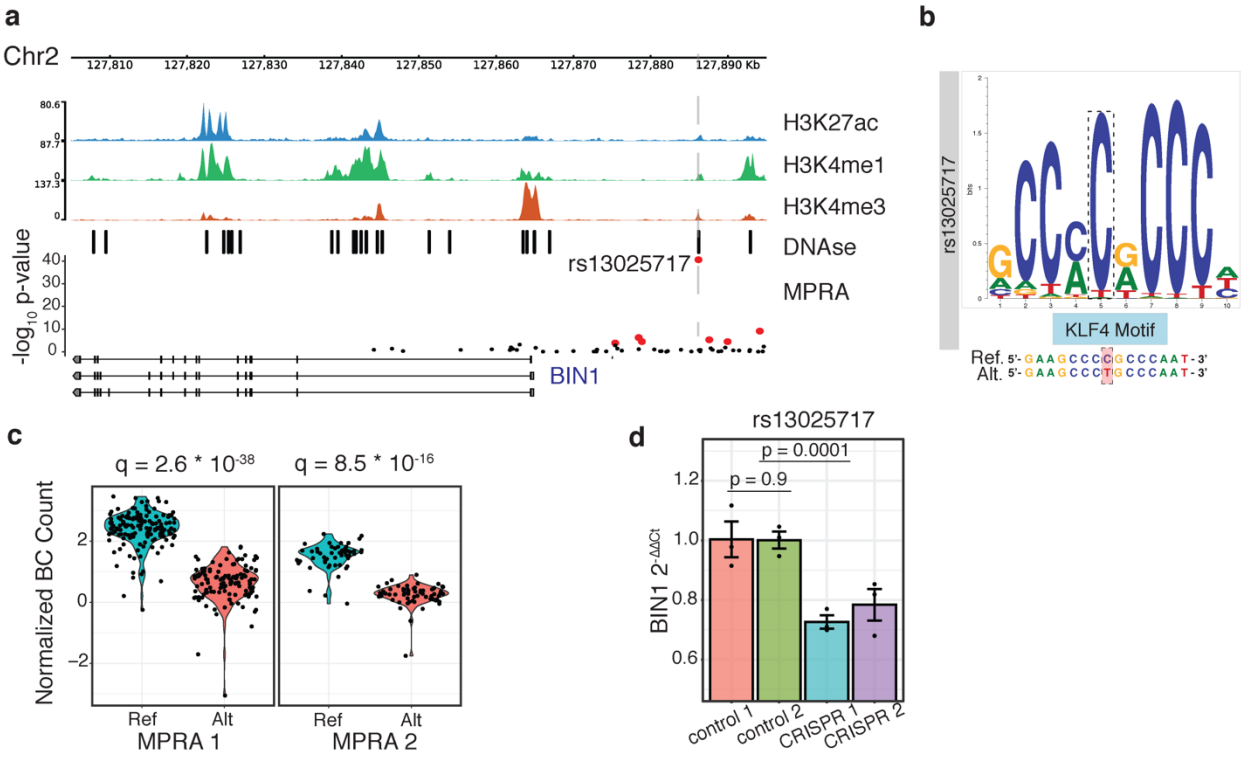


Figure 5-1. Validation of rs13025707 as a regulatory variant underlying the *BIN1* AD GWAS locus. (**A**) Genomic tracks (1-4) at the *BIN1* locus show rs13025707 falling within H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by -$\log_{10}$ p-value (track 5). SigVars (FDR q < 0.01, BH method) shown in red. (**B**) the alternate allele of rs13025707 is predicted to disrupt binding of *KLF4*. (**C**)

violin plots show normalized barcode distributions for each allele across both MPRA stages. FDR-adjusted p-values displayed. (**D**) CRISPR-mediated deletion of small genomic regions containing rs13025707 in SH-SY5Y cells significantly reduces *BIN1* expression compared with gRNA-scramble controls (n=3/group, combined n=6/condition; t(10) = -6.0; p = 0.0001; two-tailed Student's t-test). Error bars = S.E.M.

MS4A6A gene expression is regulated by rs636317

I next tested intergenic variant, rs636317, which lies within the complex Chr11:59923508 AD GWAS locus (*MS4A* gene region) [26] and is another highly significant MPRA SigVar (MPRA 1 $q = 8.7 \times 10^{-21}$). This variant is predicted to disrupt *CTCF* binding and overlaps H3K27ac and H3K4me1 peaks, as well as DHS in microglia and monocytes (Figure 5-2a-c). Analysis of existing Hi-C data from THP-1 cells [27] reveals rs636317 physically looping with upstream distal gene *MS4A6A*, which although not the closest gene to this variant, is also an eQTL [28] and the most highly-expressed gene at this locus in monocytes. As predicted, removing rs636317 in THP-1 cells, a macrophage-microglia related cell line [27] led to a significant reduction in expression of *MS4A6A* compared with both controls, validating the function of this locus in a native context (Figure 5-2d).

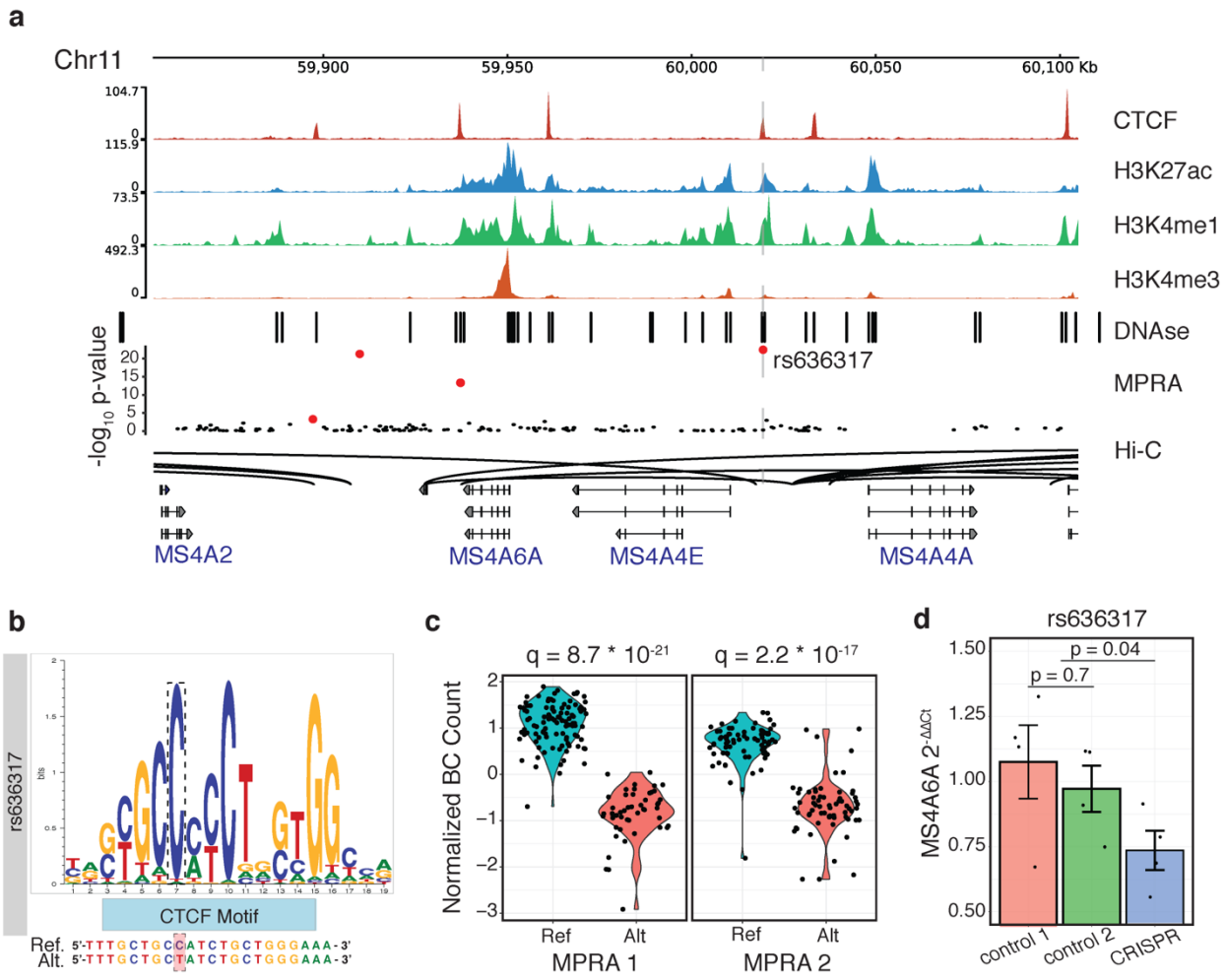


Figure 5-2. Validation of rs636317 as a regulatory variant underlying the *MS4A6* AD GWAS locus. (**A**) Genomic tracks (1-5) at the *MS4A6* locus show rs636317 falling within CTCF, H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by -$\log_{10}$ p-value (track 6). SigVars shown in red. Monocyte Hi-C data [27] (track 6) links rs636317 with *MS4A6A*. (**B**) rs636317 is predicted to disrupt binding of *CTCF*. (**C**) violin plots show normalized barcode distributions for each allele across both MPRA stages. FDR-adjusted p-values shown. (**D**) CRISPR-mediated deletion of a genomic region containing rs636317 in differentiated THP-1 monocytes significantly reduces *MS4A6A* expression (n=4/group; t(10) = -2.3; p = 0.03; two-tailed Student's t-test). Error bars = S.E.M.

CASS4 gene expression is regulated by rs6064392

The *CASS4* gene, although identified as residing within an AD-risk locus for close to a decade [26], remains relatively understudied. I identified functional variant rs6064392 within a

microglial enhancer in an intergenic region upstream of *CASS4*. This variant falls within a monocyte H3K27ac peak and is predicted to significantly disrupt ATF-family TFBS (Figure 5-3a-b). The minor T-allele of rs6064392 is in tight LD ($r^2 = 0.91$) with the protective allele of the GWAS locus lead SNP rs6024870 (A; OR = 0.88)[29] and is also predicted by MPRA to decrease downstream gene expression (MPRA $q = 2.6 \times 10^{-20}$). This is in agreement with whole blood eQTL data from the GTEx consortium (Figure 5-3c). I used a pair of gRNAs to excise a 382 bp region around rs6064392 in differentiated THP-1 cells, which significantly altered *CASS4* but not neighboring *RTF2* expression (also a predicted eQTL) (Figure 5-3e-f), supporting the functional impact of this variant on the *CASS4* gene.

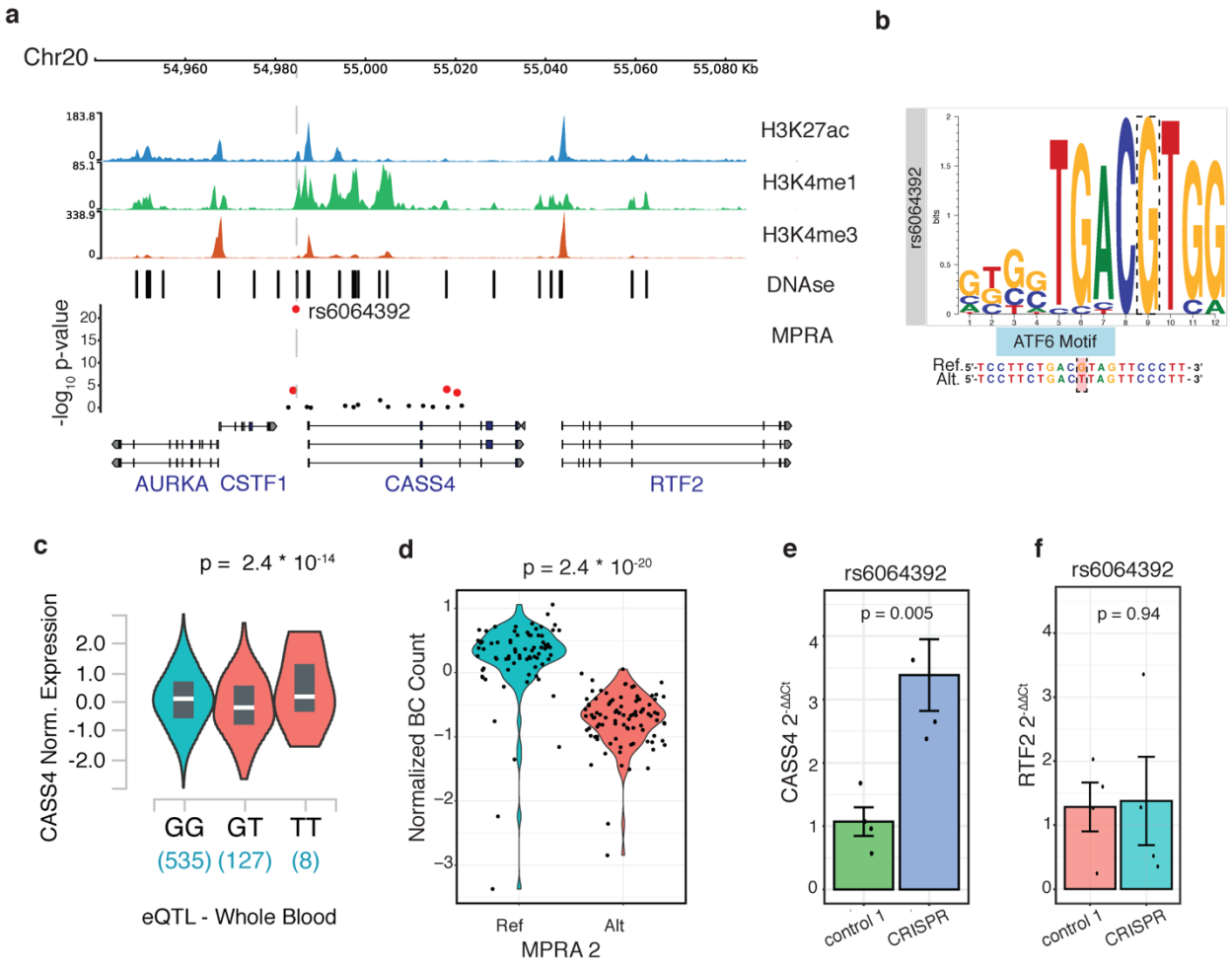


Figure 5-3. Validation of rs6064392 as a regulatory variant underlying the *CASS4* AD GWAS locus. (**A**) Genomic tracks (1-4) at the *CASS4* locus show rs6064392 falling within H3K27ac, H3K4me1, and DHS peaks (CD14+ monocytes; ENCODE) upstream of the *CASS4* TSS. All variants tested by MPRA in the locus plotted by $-\log_{10}$ p-value (track 5). SigVars (FDR $q < 0.01$, BH method) shown in red. (**B**) rs6064392 is predicted to disrupt binding of *ATF6*. (**C**) Whole blood eQTL (GTEx) for rs6064392 and *CASS4*. (**D**) violin plot shows normalized barcode distributions for each allele (MPRA 2). FDR-adjusted p-values displayed (two-tailed Mann-Whitney-U test). (**E-F**) CRISPR-mediated deletion of a small genomic region containing rs6064392 in differentiated THP-1 cells significantly reduces (**E**) *CASS4* (n=4/group; $t(6) = -4.3$; $p = 0.005$), but not (**F**) *RTF2* (n=4/group; $t(6) = 0.1$; $p = 0.94$) gene expression compared with gRNA-scramble controls. Two-tailed Student's t-test. Error bars = S.E.M.

Multiple genes regulated by rs9271171 within the HLA locus

Next, I explored the *HLA-DRB1/5* AD GWAS locus, which is a highly polymorphic region of extended LD with numerous AD-associated variants and potential risk genes. My screen identified nine significant variants in the locus, the majority of which fell within the intergenic region between *HLA-DRB1* and *HLA-DQA1.* I chose to further validate rs9271171 as it fell within a large myeloid open-chromatin region and was a significant eQTL in whole blood for multiple genes (GTEx)[28] where the alternate (protective) allele decreased reporter expression (MPRA $q = 4.9 \times 10^{-6}$). I used two pairs of gRNAs to excise part of the enhancer region containing the variant in differentiated THP-1 macrophages, which revealed a dramatic (2.5-fold; Figure 5-4b) increase in expression of *HLA-DQA1*. Interestingly, unpublished chromatin conformation data from PsychENCODE indicated that rs9271171 might also physically interact with the distal gene, complement 4 (*C4A),* for which it is also an eQTL in whole blood (GTEx)[28]. Indeed, I observed that excision of this variant significantly reduced *C4A* expression in THP-1 macrophages (Figure 5-4c). Its regulation of both *HLA-DQA1* and *C4A,* but in opposite directions, suggested that this region may be pleiotropic and regulate multiple downstream genes. Because complement components are also expressed by reactive astrocytes, I performed

the experiment again in human PSC-derived astrocyte cultures (Figure 5-4d), confirming that excision of this variant reduced *C4A* gene expression in astrocytes as well.

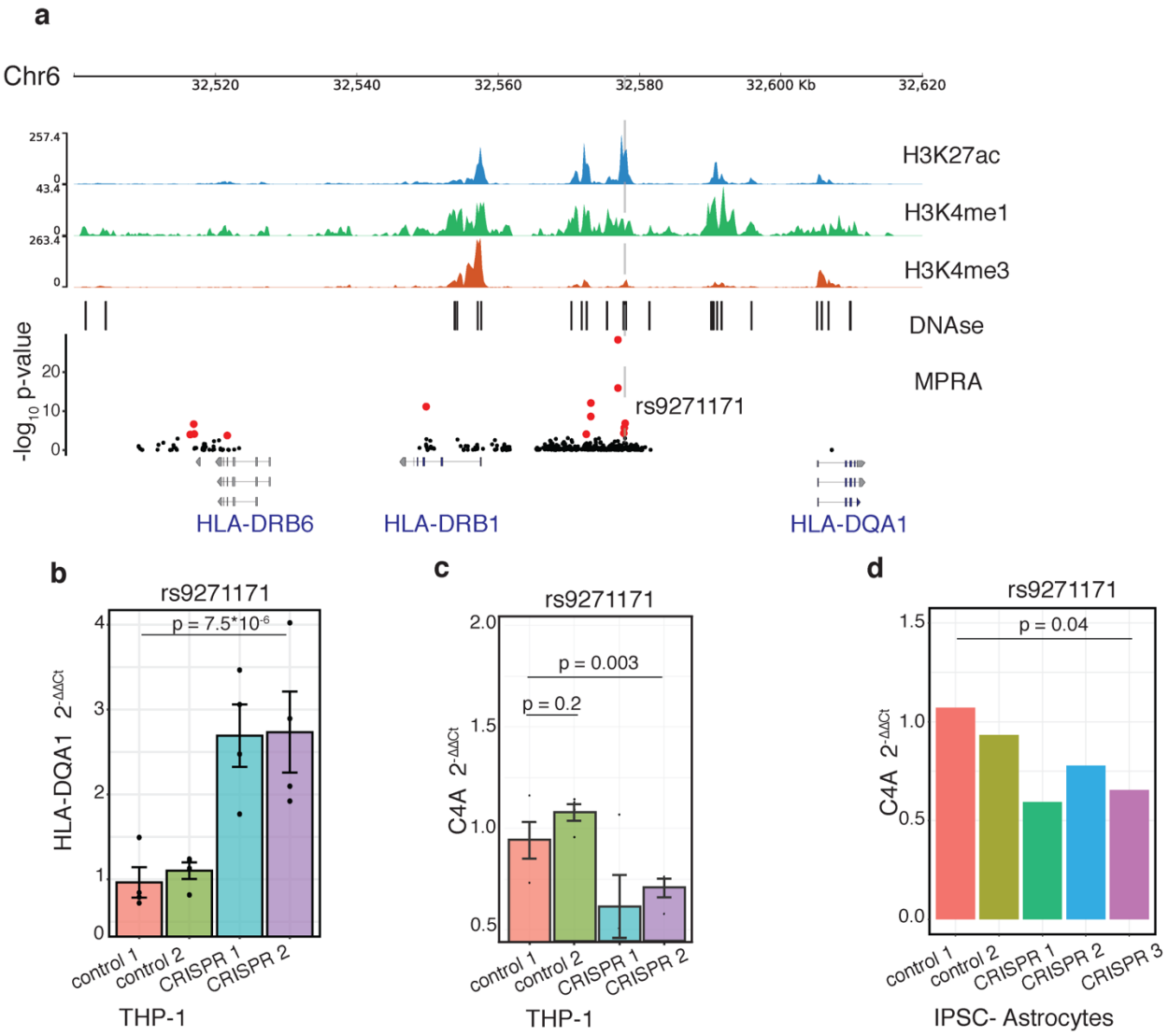


Figure 5-4. Validation of rs9271171 as a pleiotropic regulatory variant within the *HLA-DRB1/5* AD GWAS locus. (**A**) Genomic tracks (1-4) at the *HLA-DRB1/5* locus show rs9271171 falling within H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by -$\log_{10}$ p-value (track 6). SigVars shown in red. (**B**) CRISPR-mediated deletion of genomic regions containing rs9271171 significantly increases expression of *HLA-DQA1* (n=4 group, n=8/condition, t(14) = -6.9; p = 7.5×10^{-6}) and (**C**) significantly decreases *C4A* expression (n=4 group, n=8/condition; t(14) = 3.6; p = 0.003) relative to controls in differentiated THP-1 cells. (**D**) CRISPR deletion also decreases expression of *C4A* in IPSC-derived astrocytes (CRISPR n = 3, control n = 2; t(3) = 3.5; p = 0.04). Two-tailed Student's t-test. Error bars = S.E.M.

Inconclusive evidence for regulation of *CLU* gene expression by rs1532277

The clusterin (*CLU*) gene is secreted multifunctional glycoprotein implicated in amyloid clearance and inflammatory signaling, and was one of the earliest identified AD risk genes [30]. Previous studies have postulated multiple causal variants within the *CLU* GWAS locus, including rs1523378 within intron 3 of the clusterin gene [31]. My MPRA identified rs1532277 as being a significant regulatory variant (MPRA 1 $q = 4.7 \times 10^{-5}$). Of note, this variant is located just 135 bp upstream of rs1523378 within *CLU* intron 3. Furthermore, this variant falls within open chromatin in multiple cell types and is predicted to disrupt binding of the USF1 transcription factor (Supplemental Table 1). I used three pairs of guide RNAs to excise a 393, 259, or 200 bp region surrounding this variant in differentiated SH-SY5Y cells, ensuring that these truncations did not include any of the nearby exon. Confusingly, the longest excision (of 393 bp) increased expression of the secreted CLU isoform, while the shorter excisions reduced expression relative to controls (Figure 5-5a). Of note, the longer excision also deleted rs1532278, while the shorter two spared this variant from removal (Figure 5-5b).

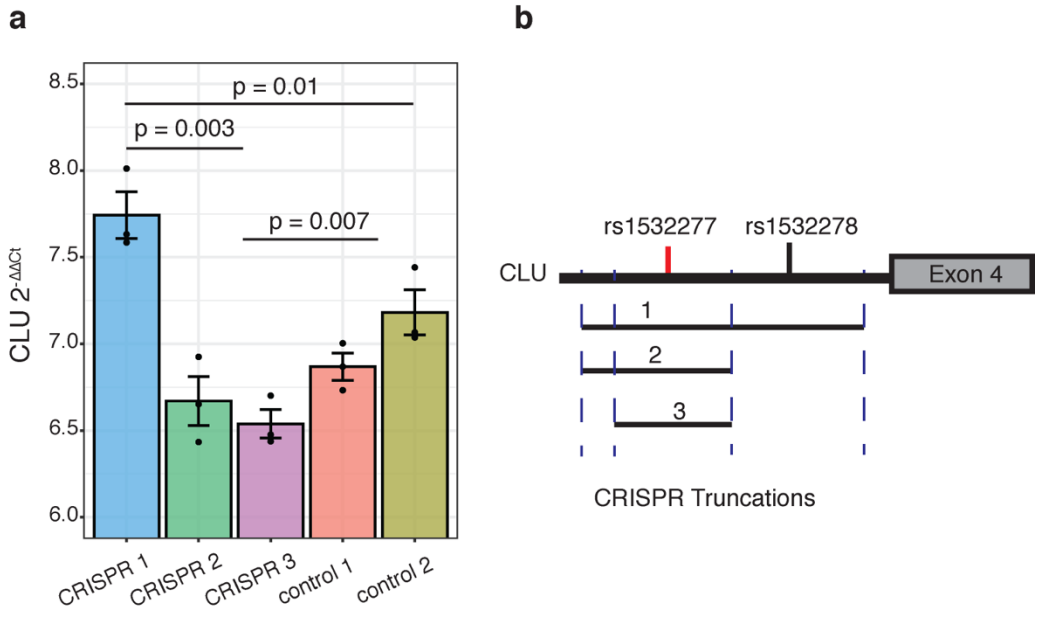


Figure 5-5. Ambiguous functional validation of rs1532277 within the *CLU* AD GWAS locus. (**A**) CRISPR mediated excision of the genomic region containing rs1532277 within intron 3 of the *CLU* gene using three pairs of guide RNAs in differentiated SH-SY5Y cells resulted in contradictory findings. The first pair (CRISPR 1) elevated gene expression (p = 0.01), while the other two pairs (CRISPR 2 and 3) reduced gene expression (p = 0.007) relative to scrambled gRNA controls. Two-sided Welch's t-test. Error bars = S.E.M. (**B**) gene model of *CLU* intron 3 and exon 4 shows the location of the genomic excisions of the three CRISPR gRNA pairs relative to rs1532277 and neighboring variant rs1532278.

rs7920721 does not regulate gene expression in THP-1 cells

Finally, the variant rs7920721 falls within an intergenic region on chromosome 10 annotated to the *ECHDC3* gene. This variant was a highly significant MPRA SigVar (MPRA q = 1.8×10^{-10}) and resides within microglial open chromatin and an H3K27ac peak. I excised the region surrounding this variant in differentiated THP-1 cells (tested because of their similarity to microglia), but failed to find any impact on gene expression for either *ECHDC3* or *USP6NL* (closest genes; Figure 5-6). However, I cannot rule out that this variant may function within a different cell type.

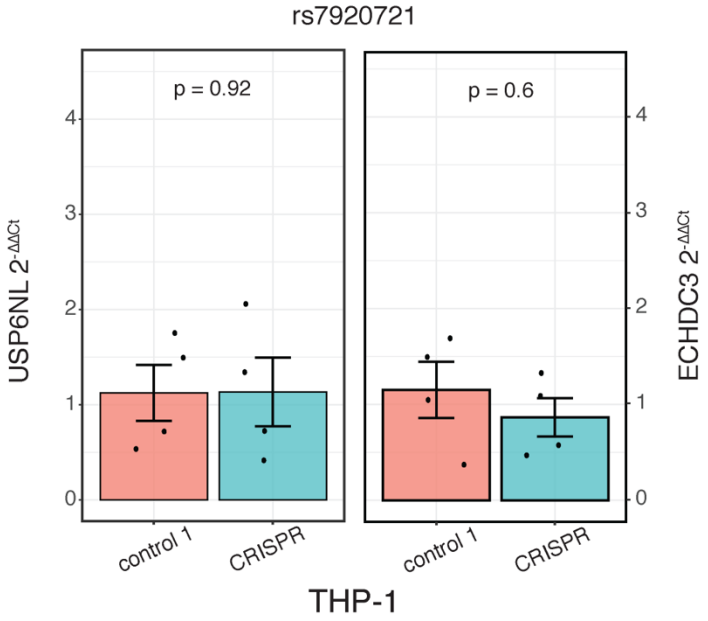


Figure 5-6. CRISPR-mediated deletion of a small genomic region containing rs7920721 in differentiated THP-1 cells does not alter expression of *USP6NL* or *ECHDC3* (n=4/group; t(6) = 0.1 and 0.6; p = 0.92 and 0.6 respectively; two-tailed Student's t-test). Error bars = S.E.M.

Discussion

I previously used MPRA to characterize noncoding variation associated with two neurodegenerative disorders, AD and PSP, identifying putative regulatory variants at numerous risk loci (Chapter 2). However, a limitation of that work was the technical infeasibility of performing MPRA within human brain cell types, with the specific *trans*-regulatory environment of HEK293T cells likely influencing the generalizability of the screen [19]. Here, I address this limitation by identifying and validating causal SNPS at six loci within brain-relevant cell lines, an important proof-of-principle for the external validity of my screen. In doing so, I provide strong evidence for rs13025717 and rs636317 as regulatory variants underlying the *BIN1* and *MS4A6A* loci respectively, as had been previously suggested [32,33]. Moreover, my work suggests that these variants likely function within microglia to promote AD risk.

Furthermore, I show that rs6064392 regulates expression of *CASS4* within monocyte/microglial lineages. The function of *CASS4* has largely been inferred by its homology with other CASS-family proteins and its interactions with known AD risk-genes such as *PTK2B/Pyk2* at focal adhesions. Interestingly, multiple studies assessing gene expression from sorted brain tissues have determined that *CASS4* is most highly expressed in microglia and endothelial cells [34,35], where it may influence cell migration, phagocytosis, and Aβ peripheral clearance [36]. I find the minor T-allele of rs6064392 is predicted by MPRA to decrease downstream gene expression, which is in agreement with whole blood eQTL data (Figure 5-3c). As rs6064392 is in tight LD with the protective A-allele of rs6024870, decreased *CASS4* expression may play a protective role in AD. This is in-line with functional genetic screens that

found elevated *CASS4* expression exacerbating tau pathology in *drosophila* models [37], although the exact role of *CASS4* in microglia remains unclear.

Furthermore, I characterized rs9271171, which lies within the flank of a recently described myeloid super-enhancer [38] in the *HLA-DRB1/5* locus. Excision of the ~600 bp surrounding rs9271171 in THP-1 macrophages dramatically increased expression of the most proximal gene, *HLA-DQA1*. Highlighting the regions regulatory complexity, in THP-1 macrophages and IPSC-derived astrocytes, deletion of this region also significantly reduced expression of complement 4 (*C4A*), a distal gene ~600 kb away. In this study, the rs9271171 (C)-allele increased expression of the reporter gene, in agreement with the eQTL effect for this allele on *C4A* (GTEx) [28]. This (C) allele is in tight LD ($r^2 = 0.99$) with the risk (A) allele of the locus lead SNP rs9271058 (OR = 1.1) [29], consistent with a hypothesis that elevated C4A expression increases risk for AD. Indeed, *C4* as well as other members of the classical complement cascade (e.g. *C1q, C3*) associate with amyloid plaques and are dramatically elevated in AD brain [39]. Work in animal models has consistently demonstrated an exacerbation of tau pathology and synaptic dysfunction downstream of complement activation [40,41]. Likewise, genetic risk in the complement pathway was identified more than a decade ago in *CR1* and *CLU* [42], in *C7* in a Han Chinese cohort more recently [43], and suggested in *C4* through a small study assessing local copy number variation [44]. However, this is the first time common genetic variation influencing *C4* expression has been linked to AD risk.

Finally, I was unable to conclusively verify the gene-regulatory effects of two SigVars, rs7920721 (*ECHDC3* locus) and rs1523377 (*CLU* locus). While excision of the genomic region containing rs7920721 failed to have any discernable effect on gene expression of nearby genes *ECHDC3* and *USP6NL*, excision of rs1523377 had contradictory results. Removal

of a relatively large genomic region containing the variant increased expression of the secreted CLU isoform, while more restricted truncations decreased gene expression. It is possible that these different length truncations include or exclude different transcriptional-regulatory domains, and indeed the longer truncation also removed nearby variant rs1532278. However, rs1532277 is an intronic variant neighboring an exon, and excision of this region might thereby impact mRNA splicing or stability, confusing the interpretability of my results. These findings highlight the limitations of my gene editing approach, wherein a small genomic region surrounding the SNP of interest is removed using pairs of guide RNAs. This method is much more technically efficient than laborious allelic replacement experiments, and provides information about whether the gene region containing the SNP of interest (in its native chromatin context) is involved in gene regulation. However, this method does not directly test the causality of that SNP due to the confounding effects of removing multiple neighboring nucleotides. Although I provide strong evidence using both MPRA and CRISPR-mediated excision, future gene-editing studies performing precise allelic replacement will be required to fully validate these likely regulatory variants.

Bibliography

1. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Cano-Gamez, E. & Trynka, G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* **11**, (2020).
3. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1–13 (2018).

4. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956 (2009).

5. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

6. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).

7. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).

8. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

9. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

10. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

11. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nat. Methods* **10**, 977–979 (2013).

12. Liu, X. S. *et al.* Editing DNA methylation in the mammalian genome. *Cell* **167**, 233–247 (2016).

13. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).

14. Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat. Commun.* **8**, 1–9 (2017).

15. Dixit, A. *et al.* Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).

16. Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).

17. Kampmann, M. CRISPR-based functional genomics for neurological disease. *Nat. Rev. Neurol.* **16**, 465–480 (2020).

18. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts. *Biol. Psychiatry* (2020).

19. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

20. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

21. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).

22. Wang, C. *et al.* Scalable production of iPSC-derived human neurons to identify tau-lowering compounds by high-content screening. *Stem Cell Rep.* **9**, 1221–1233 (2017).

23. Patel, A. M. *et al.* Dystrophin deficiency leads to dysfunctional glutamate clearance in iPSC derived astrocytes. *Transl. Psychiatry* **9**, 1–21 (2019).

24. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C T method. *Nat. Protoc.* **3**, 1101 (2008).

25. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 1–15 (2018).

26. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452 (2013).

27. Phanstiel, D. H. *et al.* Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. *Mol. Cell* **67**, 1037–1048 (2017).

28. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

29. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

30. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088 (2009).

31. Foster, E. M., Dangla-Valls, A., Lovestone, S., Ribe, E. M. & Buckley, N. J. Clusterin in Alzheimer's disease: mechanisms, genetics, and lessons from other pathologies. *Front. Neurosci.* **13**, 164 (2019).

32. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid cell genomics identifies novel causal variants, regulatory elements, genes and pathways. *bioRxiv* 694281 (2019).

33. Corces, M. R. *et al.* Single-cell epigenomic identification of inherited risk loci in Alzheimer's and Parkinson's disease. *bioRxiv* (2020).

34. Zhang, Y. *et al.* Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).

35. Srinivasan, K. *et al.* Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat. Commun.* **7**, 1–16 (2016).

36. Dourlen, P., Kilinc, D., Malmanche, N., Chapuis, J. & Lambert, J.-C. The new genetic landscape of Alzheimer's disease: from amyloid cascade to genetically driven synaptic failure hypothesis? *Acta Neuropathol. (Berl.)* 1–16 (2019).

37. Dourlen, P. *et al.* Functional screening of Alzheimer risk loci identifies PTK2B as an in vivo modulator and early marker of Tau pathology. *Mol. Psychiatry* **22**, 874–883 (2017).

38. Majumder, P. *et al.* A super enhancer controls expression and chromatin architecture within the MHC class II locus. *J. Exp. Med.* **217**, (2020).

39. Walker, D. G. & McGeer, P. L. Complement gene expression in human brain: comparison between normal and Alzheimer disease cases. *Mol. Brain Res.* **14**, 109–116 (1992).

40. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472 (2018).

41. Carpanini, S. M., Torvell, M. & Morgan, B. P. Therapeutic inhibition of the complement system in diseases of the central nervous system. *Front. Immunol.* **10**, 362 (2019).

42. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).

43. Zhang, D.-F. *et al.* Complement C7 is a novel risk gene for Alzheimer's disease in Han Chinese. *Natl. Sci. Rev.* **6**, 257–274 (2019).

44. Zorzetto, M. *et al.* Complement C4A and C4B gene copy number study in Alzheimer's disease patients. *Curr. Alzheimer Res.* **14**, 303–308 (2017).

CHAPTER 6

Technical factors influencing MPRA performance

Abstract

Massively Parallel Reporter Assays are a recently developed high-throughput experimental approach for the transcriptional regulatory characterization of noncoding genomic elements. However, this technology remains relatively immature, and the influence of key technical parameters and design considerations on assay performance remains unclear. In Part 1 of this Chapter, I use my MPRA data to explore the significance of various technical factors. After confirming that my assay is highly robust and reproducible, I determine that high MPRA barcode complexity (~40 barcodes/allele) and cellular expression is critical for assay performance. Additionally, I find modest effects of element orientation on experimental outcomes, and also determine that a library sequencing depth of approximately 25-fold maintains adequate data quality while reducing assay costs. Finally, I describe sequence features including GC content extremes, low sequence complexity, and long nucleotide repeats that predict element dropout during library construction. In Part 2, I describe technical barriers and challenges faced towards implementing MPRA within neuronal cell types relevant for modeling neurodegenerative disease.

Introduction

A major goal of modern genetics is the identification and functional annotation of noncoding genomic regions. However, characterization of the millions of putative regulatory elements identified across diverse and myriad biospecimens is a daunting task, necessitating scalable, multiplexed methods. Massively Parallel Reporter Assays (MPRA) have emerged as a high-throughput approach for functional characterization of the noncoding genome [1]. Briefly, MPRA measures the transcriptional regulatory effects of short genetic elements, using deep

sequencing to quantify transcribed barcodes uniquely associated with each assay element (Chapter 1). Since its introduction in eukaryotic systems in 2012, MPRA has been used across diverse experimental settings to screen for enhancer activity [2,3], identify functional variants [4–7], and probe transcriptional architecture [8]. However, the assay remains relatively immature, and unlike other high-throughput methodologies in functional genomics, lacks a consensus and robustly validated assay pipeline. Implementations in the field have been diverse, and the performance effects of library design, cellular environment, sequencing parameters, and analysis approaches remain to be fully clarified [9,10].

Nevertheless, there have been attempts in the literature to elucidate assay parameters critical to MPRA performance, the most comprehensive of which was a recently published study from Klein and colleagues who screened 2,440 previously described liver enhancers across nine different library designs [9]. They determined that the most significant design consideration was the spatial location of the enhancer element, which can have a canonical 5' placement or be placed within the 3'UTR of the reporter gene (as in STARR-seq [11]). The authors also found modest effects when comparing chromatin integrating (lentiMPRA [12]) vs. episomal assays. Interestingly, they further tested candidate enhancers at three different lengths (192, 354, and 678 bp), determined that the distal enhancer context influences assay results while noting that the short, sub-200 bp enhancer size seen in most studies is an artificial technical constraint of microarray synthesis [9]. In summary, Klein and colleagues systematically tested a number of MPRA library design parameters, describing advantages and disadvantages for each implementation. Importantly in regards to my work, they claim that the "classic" MPRA design based on the pGL4 reporter construct [2] (used here) had the highest reproducibility, largest dynamic range, and was the most well-predicted by regression on relevant genomic features [9].

The work by Klein and colleagues comprehensively replicates results from earlier studies characterizing episomal vs. integrating assays [12,13] and 5' vs. 3' enhancer placements (STARR-seq [11]). Other parameters examined in the literature include power at different barcode complexities [5,14] or numbers of biological replicates [15], influence of minimal promoter strength [16–18], impact of library expression and coverage [19], or the effects of different modeling assumptions on data analysis pipelines [15,20–22]. Furthermore, it has been extensively noted that cellular context and specific *trans*-factor expression profiles dramatically influence reporter assays results, and several studies have examined cell-type specific effects [5,8,16,23] .

In this work I use the MPRA data I previously generated (Chapter 2) to examine technical factors influencing MPRA performance. In Part 1, I discuss the effects of a number of technical factors including variant dropout during library construction, barcode complexity, library expression, and sequencing depth. These results will inform the discussion in Part 2, where I note the significance of cellular context and describe my progress in adapting my MPRA approach for use in neuronal cells.

Materials and Methods

Data and quality control

MPRA data used and visualized here was previously generated and described (Chapter 2). Quality control (inter-replicate and inter-assay) were also previously defined.

Variant dropout analysis

Variants were excluded from analysis if I was unable to obtain activity measurements from at least 5 unique barcodes for both alleles. Of the 5,706 unique SNPs tested in both MPRA

stages, 366 did not meet this inclusion threshold in either stage. I then compared sequence features from these "dropouts" vs the rest of the "included" oligos. I used the SeqComplex perl module to compute GC content, CpG skew, and sequence complexity metrics (https://github.com/caballero/SeqComplex) [24]. Sequence entropy was calculated using the RNAfold tool from the ViennaRNA 2.0 software suite [25]. I also generated a custom python script to identify the longest runs of mono- or di-nucleotide repeats within each oligo (script provided: https://github.com/ycooper27/Tauopathy-MPRA). Outputted scores were compared between "dropouts" and all "tested" oligos using a two-sided Mann-Whitney-U test. For visualization purposes (Figure 6-3) I took a random sample of 366 "tested" variants.

Power analysis

I performed a power analysis to determine the sensitivity of my assay at different barcode complexities. I first determined empirical SigVar effect sizes from my combined study, which were binned into percentiles. I also determined an empirical assay standard deviation by taking the average standard deviation of the normalized barcode counts for all alleles that passed filter in both studies. I then performed a power analysis using the power.t.test function from the stats package in R, using the empirical standard deviation, an alpha threshold of $p < 0.01$ and a "two.sided" hypothesis test. The analysis was performed using different percentiles (0, 20, 40, 60, 80th) of empirically computed effect sizes, as well as all integer n (i.e. # unique barcodes per allele) between 5-100, and plotted using ggplot2.

AAV production

The pAAV.CMV.PI.EGFP.WPRE (Addgene #105530) plasmid was packaged into AAVs of four different serotypes using the following rep/cap plasmids: pAAV2/1 (Addgene #112862) , pAAV2/9n (Addgene #112865), 7m8 (Addgene #64839) [26], and pUCmini-iCAP-PHP.eB (Addgene #103005) [27]. This was done using the triple transfection method and a modified iodixanol gradient purification protocol [28]. In brief 1:1:1 molar ratios of insert, pAdDeltaF6 helper plasmid (Addgene #112867), and rep/cap plasmid were transfected into eight 15 cm plates of 70% confluent HEK293T cells using Lipofectamine 2000, following manufacturer's instructions. Six hours post transfection the media was switched to low (2%) serum culture media. Cells were mechanically detached 72 hours later, spun down, and lysed via three freeze-thaw cycles using a dry ice slurry and 37°C water bath. Cell lysate was spun at 1000g for 15 min and the liquid phase was subsequently loaded into an 54/40/25/15% underlayed iodixanol gradient in a Beckman Coulter tube (Beckman 331372) and spun for 5 hours at 35,000 rpm and 4°C using a SW41 rotor and a swing-bucket ultracentrifuge. 3 mL at the 54/40% interface were removed and loaded into an Amicon concentrator (Millipore UFC91008) for buffer exchange and concentration in DPBS + 0.001% Pluronic-F68. The MPRA 1 library (Chapter 2) was also packaged into PHP.eB serotype AAVs using this method. Viral titer was determined by qPCR using primers for the eGFP insert and a DNA standard.

AAV serotype comparison

The pAAV.CMV.PI.EGFP.WPRE plasmid was packaged into AAV 1, 9, 7m8 [26], and PHP.eB [27] serotypes as described above and viral titer was determined by qPCR. Human neural progenitor cells (NPCs) were expanded and cultured as previously described [29]. 10^5 NPCs were

seeded onto acid-washed glass coverslips pre-coated with PLO/Laminin in 24-well culture vessels. Each AAV serotype was added in triplicate at a concentration of approximately 10^4 vg/cell, and cells were incubated for an additional five days post infection. Coverslips were then washed and fixed with 4% PFA at room temperature for 12 minutes before DAPI counterstaining and mounting onto coverslips. Slides were imaged using a Zeiss fluorescent microscope with a 20X air objective, and the proportion of GFP+/DAPI+ cells was manually quantified.

Cell culture

HEK293T cells: HEK293T cells were cultured in DMEM containing GlutaMAX (Thermo Fisher Scientific, 10566016) supplemented with 10% FBS and 1% Sodium Pyruvate (11360070).

IPSC-derived neurons: Induced Pluripotent Stem Cells (IPSCs) were previously documented [30] and kindly provided by Dr. Li Gan in accordance with the UCLA TDG guidelines. IPSCs were grown on Matrigel-coated culture vessels using MTESR media as previously described. To perform neuronal differentiation, IPSCs were gently detached using Accutase and plated on culture vessels pre-coated with Poly-L-Ornithine and Laminin. IPSC-media was switched to a differentiation media composed of: DMEM/F-12, 1X N2 supplement (Thermo Fisher Scientific, 17502048), 1% MEM-NEAA (Thermo Fisher Scientific, 11140050), Y-27632 (StemCell Technologies), Doxycycline (2ug/mL), and Laminin (200 ng/mL). On differentiation day 3, media was switched to 50% DMEM/F-12 + 50% Neurobasal A, 0.5X N2, 1X B27 (Thermo Fisher Scientific, 12587010) with the following supplements: 10ng/ml BDNF, GDNF, NT3, Laminin (200 ng/ml), RepSox (7.5uM), Doxycycline (2ug/mL), and ascorbic acid (200 nM).

On day 7, a half media change was performed (without Dox and RepSox). Half media change was performed every 3-4 days until collection.

Massively Parallel Reporter Assay

Three technical replicates of 4-5 million IPSCs were grown and differentiated as above on pre-coated 10 cm culture dishes. On differentiation day 14, a half media change was accompanied by 100 uL of MPRA 1 library packaged into PHP.eB serotype AAVs (titer ~6 x 10^{-12} vg/mL). The virus was allowed to express for seven days. Cells were collected on day 21 and DNA and RNA were extracted. Libraries were prepped for sequencing and downstream analysis as previously described (Chapter 2, Methods), with 18 PCR cycles required for cDNA and Plasmid amplification prior to sequencing.

Statistical analysis

All statistical analysis was performed using the stats package in R (v. 4.0.0). All statistical tests are reported where relevant and are two-sided.

Results

Part 1

In Chapter 2 I described screening 5,706 unique noncoding variants across two unique MPRAs. Here, I use these data to provide further discussion and analysis of factors observed to impact MPRA experimental outcomes and technical performance. Consideration of these parameters may aid in the design and implementation of future studies involving complex library construction and massively multiplexed assays.

ASSESSMENT OF MPRA REPRODUCIBILITY

Because of my two-staged experimental design, I took the opportunity to characterize the true reproducibility of my assay. I identified 326 variants in the first MPRA stage, including 186 with significant transcriptional skew between alleles (SigVars; FDR adjusted $q < 0.01$) to be re-tested in the second stage. Importantly, although library design stayed identical, oligo barcoding, cell culture, sequencing, and other technical factors were distinct between stages, thereby providing an informative estimate of the impact of technical confounding variables. Fortunately, I found that the assay was highly reproducible, with measurements of allele transcriptional efficacy, and relatedly, MPRA-determined effect sizes, highly correlated between experiments ($r = 0.98$ and 0.94 respectively, both $p < 2\times10^{-16}$; Figure 2-8a). This suggests that MPRA-determined measurements of transcriptional efficacy are highly precise and that variant ordering (e.g. prioritization) on the basis of effect size is robust. Significantly, reproducibility is not an inherent feature of SigVars or highly expressed variants in particular. Upon inspection of the 140 re-tested variants without significant allelic skew ($q > 0.01$), I again confirmed high correlation between stages ($r = 0.96$, $p < 2\times10^{-16}$)

Interestingly, 152 of 186 (82%) re-tested MPRA 1 SigVars remained significant in stage 2 (replication threshold; Bonferroni $q < 0.05$; Figure 2-8a), which is somewhat lower than might be expected considering the original conservative FDR threshold of $q < 0.01$ used to determine significance in stage 1 and the near perfect correlation of effect sizes observed between experiments. In comparing the 34 non-replicating (NR) and 152 replicating (R) variants, the non-replicates had a *higher* MPRA stage 1 median barcode complexity (average least complex allele; 94 (NR) vs. 67 (R) barcodes), and a *lower* MPRA stage 2 barcode complexity (45 (NR) vs. 65

(R) barcodes). Thus, variants that failed to replicate had a dramatic reduction in barcode complexity in the replication stage (delta barcodes = -49 (NR) vs. -3 (R) barcodes). Concurrently, non-replicating variants had a significantly lower mean absolute effect size (MPRA stage 1 Log_2 FC: 0.36 (NR) vs. 0.72 (R); $p = 2.6 \times 10^{-9}$; two-sided Mann-Whitney-U test).

I next determined power to detect significant allelic skew for a range of percentiles of empirical MPRA effect sizes at different levels of barcode complexity (Figure 6-1; Methods). Interestingly, the threshold at which 80% power is achieved for the 40^{th} percentile of empirical effect sizes is at a barcode complexity of 42 barcodes per allele, with a steep drop-off in power at decreasing effect sizes and barcode complexities. Thus power to detect SigVars of small effects are highly contingent on high (>40 barcodes/allele) barcode complexity, and suggests that while MPRA is highly precise in determining allele effect sizes (even at low complexities ~5 barcodes, data not shown), identification of allelic skew *significance* will be somewhat impacted by stochastic fluctuations in barcode complexity due to PCR. Interestingly, while barcode complexity seems that have a strong impact on sensitivity, it may play less of a role in determining specificity. For example, I did not find that increasing the stringency threshold for variant inclusion from a minimum barcode complexity of 5 to either 8 or 10 improved the reproducibility rate. Indeed, there was no obvious bias towards having low MPRA 1 barcode complexity amongst variants that failed to replicate in MPRA 2. It therefore seems likely that other factors that influence assay quality, including library expression and coverage (discussed below), may influence noise and specificity more than barcode complexity.

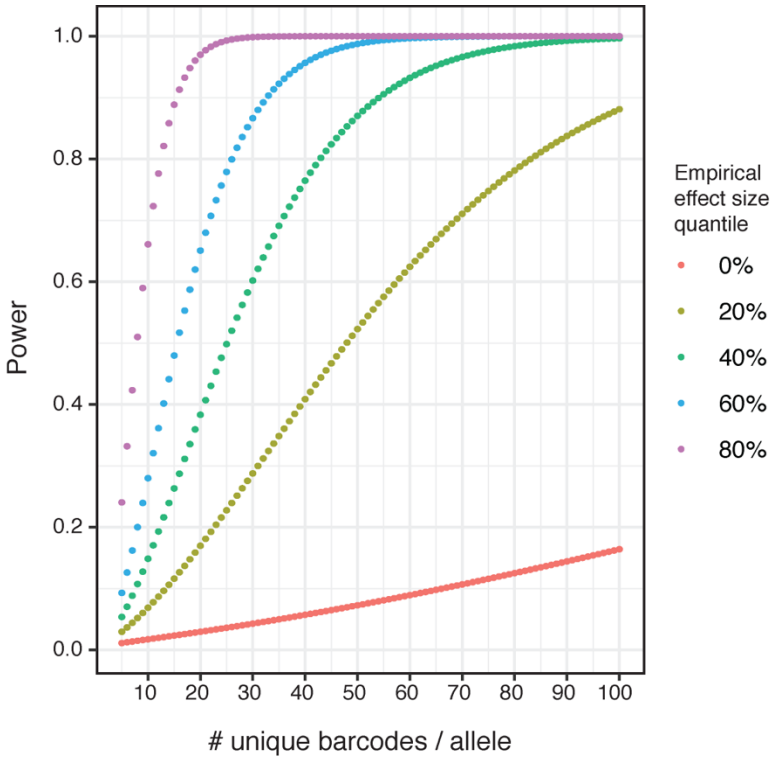


Figure 6-1. Barcode complexity significantly impacts power for massively parallel reporter assays. A simulated power analysis, using MPRA-determined empirical variance and different quintiles of empirical effect sizes, revealed that adequate power (0.8) to detect significant allelic skew at a broad range of effect-sizes is achieved at a barcode complexity of ~40 barcodes/allele.

EFFECTS OF OLIGO CONFIGURATION ON MPRA PERFORMANCE

I also tested the effects of library construction parameters on SigVar detection and reproducibility, using the SigVars identified in MPRA stage 1 as a "gold standard" test set. Specifically, I tested the impact of placing the oligo in the reverse complement (RC) orientation in the MPRA vector as well as the effect of placing the variant in the bottom third (as opposed to the middle) of the 162 bp genomic context. I used the oligo in reverse orientation as a negative control as this maintained identical nucleotide composition (Figure 6-2). Interestingly, SigVars had strong but imperfect correlation with their RC ($r = 0.69$, $p < 2\times10^{-16}$) and lower third ($r = 0.78$, $p < 2\times10^{-16}$) counterparts, suggesting modest effects of oligo orientation and distal sequence context on MPRA activity. As MPRA activity measures are highly precise and

reproducible (discussed above), these effects of oligo configuration are likely biological and not due underlying assay noise. Placing oligos in the reverse orientation completely abolished activity as expected (all $p > 0.5$; Figure 6-2). These findings can inform design considerations for future massively parallel screens.

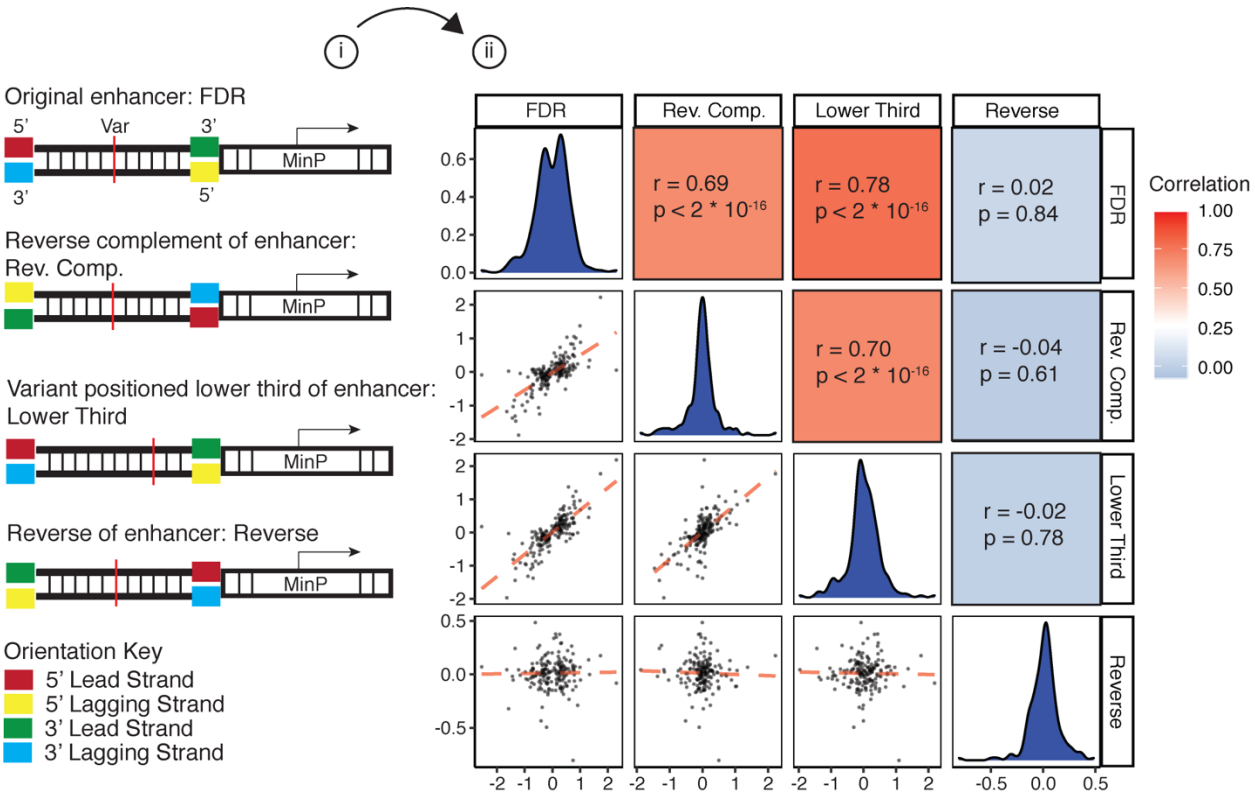


Figure 6-2. Effects of oligo configuration on MPRA performance. i) Shows the experimental design for testing the effect of enhancer orientation and variant placement on MPRA activity. 212 variants identified from MPRA 1 were replicated in MPRA 2: Oligos were kept in the same orientation as the original (FDR), oligos were placed in the reverse complement orientation (Rev. Comp.), variant was placed in the lower third of the oligo rather than the middle (Lower Third), or the enhancer sequence was reversed (Reverse) as a negative control. Plots showing correlations between oligo orientations shown in ii), red line = OLS regression line of best fit, Pearson's correlation.

TECHNICAL FACTORS INFLUENCING VARIANT DROP-OUT

I initially assumed that the observed 9% variant drop-out during stage 1 was random bottlenecking due to library construction and therefore attempted to re-test missing variants in

stage 2. However 346/491 of these variants also failed to pass QC in stage 2, suggesting that a large proportion of drop-outs were due to systematic amplification failure during library construction or strong transcriptional repression. Indeed, the vast majority of variants that dropped out specifically during MPRA 1 library construction dropped out again during MPRA 2 library construction, suggesting PCR amplification failure. I assessed sequence features of the 366 unique variants that failed to pass quality control (QC) for both MPRA 1 and 2 and found that missing variants were more likely to contain GC content greater than 75% or less than 25% (GC extremes) and had increased rates of CpG skew ($p = 8*10^{-5}$; Mann-Whitney-U test). Additionally, drop-out variants had significantly lower mean sequence complexity (measured by Shannon's entropy, compressibility, linguistic complexity measures, *etc.*; Methods), and were more likely to contain long regions of repetitive sequences such as CpGs, dinucleotide repeats, or polyA tracts (average longest repeat 14 vs. 7 nucleotides; $p < 2 *10^{-16}$; Mann-Whitney-U test). Results are summarized in Figure 6-3. The two features most predictive of variant drop-out were GC extremes and di-nucleotide repeats of more than 20 nucleotides.

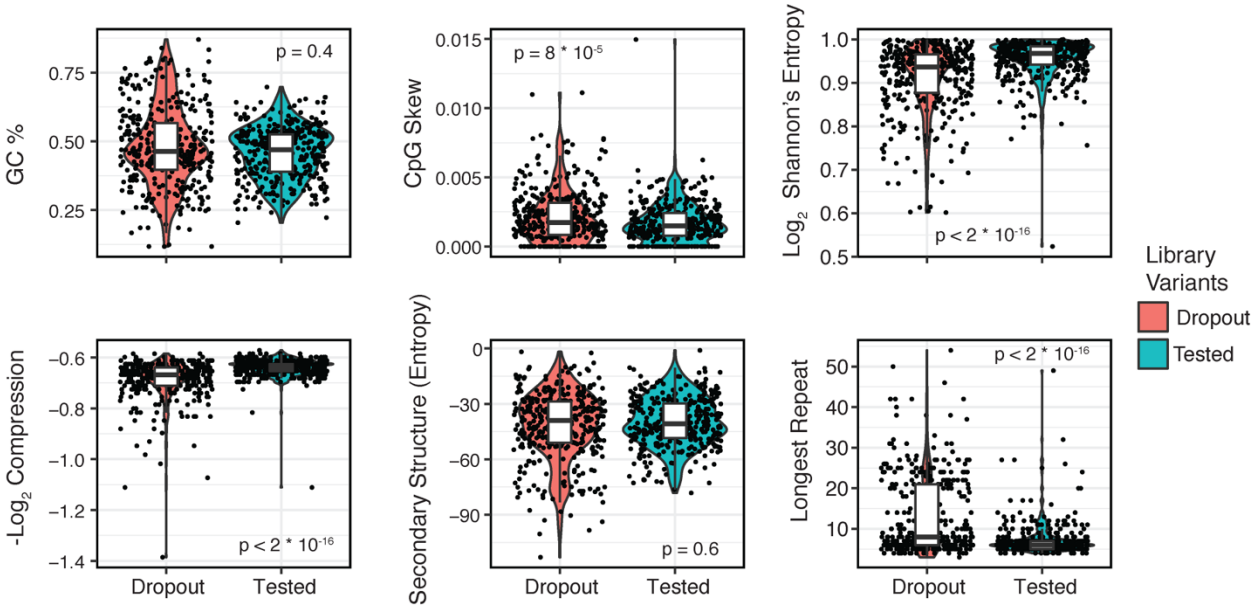


Figure 6-3. Assessment of sequence-level features of the 366 variants ("Dropouts") that failed to pass quality control thresholds through both MPRA experimental stages (Methods). Violin plots display GC-content, CpG content, sequence complexity measures ($-\log_2$ compressibility and $\log_2$ Shannon's entropy), predicted secondary structure (calculated using RNAfold), and longest mono- or di-nucleotide repeat per tested oligo for the 366 dropout variants vs. a random sample of 366 variants that passed QC ("Tested"). Dropout variants were enriched for GC content extremes (>75% or <25%), increased CpG skew, had decreased mean sequence complexity, and on average contained longer runs of nucleotide repeats (mean 14 vs 7 nucleotides; two-tailed Mann-Whitney-U test). Predicted secondary structure did not differ between groups ($p = 0.6$).

SEQUENCING DEPTH REQUIREMENTS

The appropriate sequencing depth (defined here as mean mapped reads per unique barcode) to obtain high-quality MPRA data has remained unaddressed in the literature. As sequencing is one of the largest cost-components of these assays, it is valuable to identify the optimal balance between coverage and data quality. I initially sequenced my libraries at a high, 45-fold coverage. I therefore performed read downsampling, assessing the number of unique barcodes retained and variant level inter-replicate correlations (Spearman's rho) at 45, 25, 20, 15, and 10x library coverage. As can be seen in Figure 6-4, library barcode complexity drops off down to 72% of maximum at 10x coverage, with an inflection at around 25x. Allele level inter-replicate correlation showed a similar trend (though admittedly maintaining decent correlations at all depths). These results suggest that a target sequencing depth of 20-25x should maintain appropriate data quality while reducing cost.

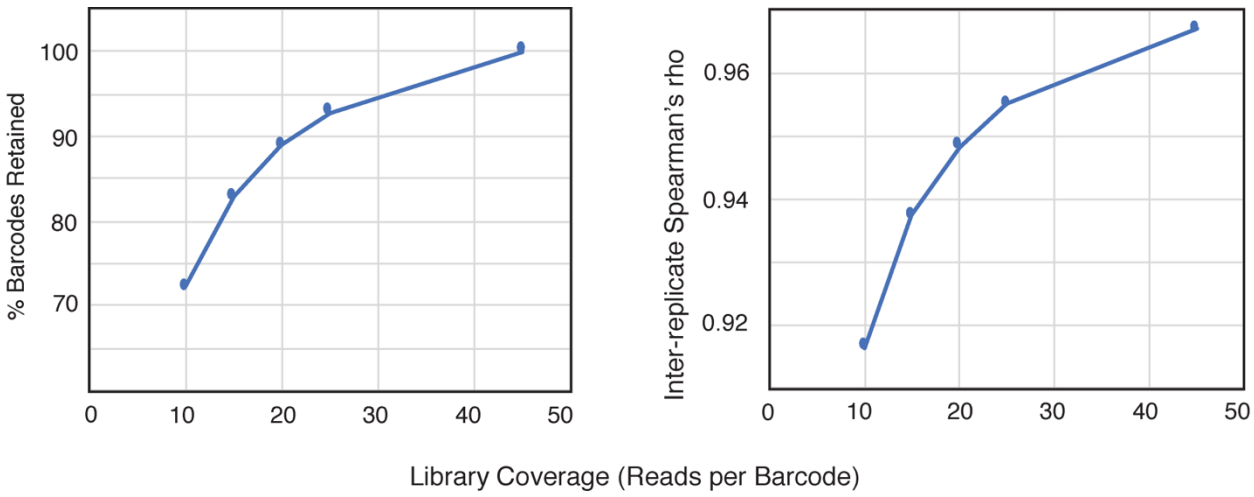


Figure 6-4. MPRA performance metrics at various sequencing depths. Library coverage was downsampled (coverage defined as the average number of reads per unique library barcode). The percentage of retained barcodes (left) and the mean variant inter-replicate Spearman's correlation (right) are shown at different read depths.

One quality control metric used in next generation sequencing studies is the percentage of reads mapped to the reference, with a target of ~90% for RNA-seq [31]. However in my MPRAs I typically observed between ~40-50% reads mapped, highlighting that a majority of reads are wasted. The algorithm I used to map reads requires perfect matching between sequenced barcodes and the reference. Therefore, errors introduced by cDNA amplification or during sequencing may account for some proportion of unmapped reads. Indeed, I found that the percentage of reads mapped to plasmid libraries was 3% higher than reads mapped to mRNA libraries, a discrepancy likely reflecting errors introduced during cDNA synthesis. I also considered another explanation: Because barcodes are appended to library elements using PCR, I use paired-end sequencing to create a reference file mapping unique barcodes to library elements (See Chapter 2, Methods). Reads are later aligned to this barcode lookup table. If this initial barcode-mapping step was not performed at sufficient coverage, it is possible that I sampled only a small proportion of true barcodes in the library. This would create a limited reference, leading to poor read mapping. Therefore, for one of my MPRA libraries, I went back and re-sequenced

barcode-variant associations at 6-fold the initial depth, identifying approximately twice the original number of uniquely mapped barcodes. However, this only improved read mapping during the actual MPRA by a few percent, suggesting that barcode saturation is not required.

The most likely explanation for poor mapability has to do with the synthesis and cloning of the library elements themselves. When assessing allelic effects, the only difference between the reference and alternate alleles can be the target SNP. Therefore, I only consider barcodes mapped to "perfect" library elements with no off-target deletions or mutations. Agilent high fidelity library synthesis advertises an error rate of one per kilobase (likely optimistic), which has a 15% probability of occurring at least once in an 162 bp element. Moreover, microarray synthesis is known to introduce frequent deletions at longer oligo lengths, and while attenuated using emulsion PCR with high-fidelity polymerase, amplification of repetitive genomic libraries can lead to a large proportion of chimeric amplicons. Indeed, I found using colony PCR that between 15-20% of library plasmids contained elements of an inappropriate size. In summary, I estimate that the process of element synthesis and cloning introduces approximately ~40% "junk" into my library. Reads from barcodes associated with these junk elements cannot be mapped, explaining the low overall percentage of reads mapped. Unfortunately, this is a difficult issue to rectify at present. However, it must be accounted for when estimating adequate assay sequencing depth.

LIBRARY EXPRESSION IMPACTS ASSAY PERFORMANCE

It has been previously reported that MPRA performance, as defined by the precision and reproducibility of measurements of library transcriptional efficacy, is dependent on high levels of library expression [19]. This is particularly true for assays measuring variant function, which

attempt to identify subtle changes in transcriptional efficacy between alleles (i.e. detect small effect size events). Three critical factors influencing expression of reporter libraries are: 1) the ability of library elements to drive transcription, 2) global transcriptional propensity of the particular cell-type, 3) transfection or infection efficiency into cells of interest (e.g. copy number per cell, cell transduction percentage). Unfortunately, assays testing variant allelic skew typically employ weak minimal promoters so as not to overwhelm the transcriptional impacts of single nucleotide substitutions, which are typically small. As a result, these assays have been exclusively performed in cancerous cell lines, which can be transfected at high efficiencies. Moreover, cancer cells, particularly those with c-Myc amplifications have been noted to have higher levels of overall transcription (known as transcriptional amplification) [32].

I confirmed the importance of library expression on assay performance by comparing four separate MPRA experiments: the two assays presented Chapter 2, the AAV-MPRA experiment (described below), as well as an additional unpublished assay. I took the average between-replicate correlation (Spearman's rho) of allele expression as the overall performance metric for each experiment. Significantly, each experiment required a differing number of PCR cycles to amplify an adequate amount of library cDNA for downstream sequencing (10, 15, 17, 18 total PCR cycles). I found an inverse relationship between PCR cycle number and performance (mean rho), with a large performance drop off above 15 cycles. As PCR cycle number likely reflects underlying mRNA quantity and library expression, this confirms the importance of expression on performance. This observation represents a major technical obstacle towards implementing MPRAs within difficult to transduce primary cell-lines, which may be desirable *in vitro* model systems that more closely recapitulate relevant biology than cancer lines (discussed further in Part 2).

MPRA DATA ANALYSIS CONSIDERATIONS

The development of analysis methods for MPRA data remains an active area of ongoing research. Previous studies examining allelic effects have used diverse approaches, including adoption of differential expression packages from RNA-seq (e.g. DESEQ2), t-tests, and non-parametric tests. More recently, a number of dedicated methods have been developed specifically tailored to MPRA data analysis, including QUASAR-MPRA [20], MPRAnalyze [21], MPRALM [15], and atMPRA [22], with user-friendly implementations in the R computing environment. I did not rigorously benchmark these various methods as this has been done extensively through simulation elsewhere [15], but I will highlight some important considerations.

A primary application of MPRA is to identify noncoding elements with enhancer or repressor activity, which is done by comparing the transcriptional activity of each library element compared with a null distribution. This distribution has been constructed in previous studies by including negative control elements such as: an "empty" vector containing only the minimal promoter, sets of scrambled sequences, or previously defined negative/control elements [1]. It is my view that the empty vector approach is dubious as there will always be some plasmid DNA elements adjacent to the minimal promoter, which may have transcriptional activity. Indeed, it has been found that the plasmid ORI can drive transcription in mammalian systems [33], as can the AAV ITR [34] (I mitigate this by introducing a transcriptional pause sequence upstream of the minimal promoter in my vector design [28]). In my work, I did not include negative control elements, and therefore defined activity as a significant deviation from the activity of the median library element. This is reasonable under the assumption that most surveyed noncoding elements are transcriptionally neutral, but may not hold for all libraries. What is certainly inappropriate is taking the normalized RNA/DNA ratio and defining active or repressed elements at thresholds

greater or less than one respectively, as these unadjusted ratios are certainly subject to sampling bias during sequencing [35] . Additionally, because I use a weak minimal promoter with low basal transcriptional activity, it is much easier to identify elements that increase rather than decrease transcription (as has been noted [5]). Repressor screens should incorporate stronger minimal promoters to increase sensitivity.

To identify allelic skew, I adapted a pipeline from Ulirsch and colleagues [4], using the non-parametric Mann-Whitney-U test to compare barcode rank distributions between alleles. Non-parametric tests are robust to deviations from normality and outliers at the expense of power. I confirmed that this approach adequately controls p-value inflation by visual comparison of Q-Q plots before and after removal of variants with $p < 0.01$. In this method, barcode counts are combined across replicates before comparison of the barcode distribution between alleles. It is therefore highly sensitive to barcode complexity (as described above). By contrast, recently developed methods for dedicated MPRA analysis are based on linear models, including the eponymous MPRALM [15]. This method explicitly models the variance-mean relationship inherent in MPRA data, and pools variance across elements using Bayesian shrinkage to increase discovery power [36]. More generally, linear models are attractive because they are fast and easy to implement, simple to interpret, and can flexibly incorporate various design matrices and contrasts to explore interaction terms. One difference is that these approaches aggregate barcodes within replicates, and the choice of aggregation method (e.g. sum, mean, median) may affect outcomes [15]. Additionally, these methods may be less sensitive to barcode complexity, but more sensitive to the number of biological replicates [15].

Part 2

A theoretical advantage of MPRA is the ability to interrogate genetic elements in tissues or cell types relevant to traits of interest [1]. Examination of eQTLs reveals that single genetic variants can simultaneously increase or decrease downstream gene expression depending on the specific tissues examined [37]. Similarly, it has been noted that variants tested in reporter assays can have opposite effects depending on the cell type used [38]. Thus, measuring variant-expression effects depends on the interaction between the variant and the specific *trans*-cellular environment, consisting of transcription factors, RNA binding proteins, and transcriptional machinery [1]. Neurodegeneration is a brain phenotype, mechanistically implicating all major brain cell types including neurons, astrocytes, oligodendrocytes, microglia, and endothelial cells. As discussed in Chapter 2 (Figure 2-3), these represent highly non-overlapping cell types and I therefore chose to use HEK293T cells as a technically tractable compromise. While ideally a functional genetic screen of neurodegeneration would occur in brain cell types, these primary cells are difficult to isolate, culture, and most importantly transduce. As I note in part one, a key technical parameter governing MPRA performance is strong library coverage and resultant expression, which is mediated by efficient cellular transduction. Here, I will describe attempts to perform MPRA within primary human neuronal model systems. Neurons were chosen as the cell type most directly mediating relevant behavioral phenotypes in neurodegeneration, and as harboring the most genetic risk for PSP [39,40] (Chapter 2).

AAV-MPRA

The implementation of MPRA within neurons is primarily limited by poor transfection efficiency. Adeno-Associated Viruses (AAV) are a commonly used viral delivery system, with well-documented neural tropism and high cellular copy numbers of the genetic payload [41].

I therefore chose to test AAVs as a library delivery system in neurons. Different AAV serotypes have different tissue tropisms and cell-type specific transduction efficiencies. First, I screened four viral serotypes, AAV1, AAV9, 7m8 [26], and PHP.eB [27] for infectivity in cultured human neural progenitor cells (hNPCs). To do so, I packaged a reporter vector expressing eGFP under control of a constitutive promoter into AAVs of these four serotypes. I then infected culture wells containing hNPCs with equal amounts of each viral serotype and manually determined the percent of GFP+ cells per field using fluorescent microscopy. I found that AAV-PHP.eB seemed to have the best tissue tropism and it was selected for the subsequent experiment.

I previously discussed the development of an MPRA vector compatible with AAV packaging and delivery (Chapter 2, Methods). Since my MPRA library plasmids were designed such that functional elements were harbored between the AAV2 ITRs, I was able to directly package my library into PHP.eB serotype AAVs using an in-house protocol (Methods). I chose to use neurons derived from induced pluripotent stem cell (IPSC) using the NGN2-overexpression protocol as this differentiation paradigm delivers rapid and homogenous neuronal differentiation [30]. I infected three biological replicates of 5 million IPSC-derived neurons, allowed my library to express for seven days, collected mRNA and DNA and sequenced the library.

However, data quality was quite poor. The log-normalized barcode count distribution exhibited a large peak at very high counts (Figure 6-5a), indicating the presence of PCR-overamplification artifacts comprising a large proportion of the data. Indeed, this library required 18 PCR cycles to obtain barely enough material for sequencing, which is many more amplification cycles than needed in HEK293T cells. Moreover, I found low and highly non-linear barcode level correlations between replicates (mean Pearson's $r = 0.49$, $p = 2*10^{-16}$).

Variant level inter-replicate correlations had a mean Pearson's correlation of 0.52 (Figure 6-5b), which is dramatically lower than the correlations of 0.95-0.99 observed in HEK293T cells. Similarly, when comparing variant expression and allelic effect size z-scores between iNeurons and HEK293T cells, there was almost no correlation (Spearman's rho 0.06-0.17; Figure 6-5c). These results suggest that the AAV-MPRA protocol failed to deliver usable data.

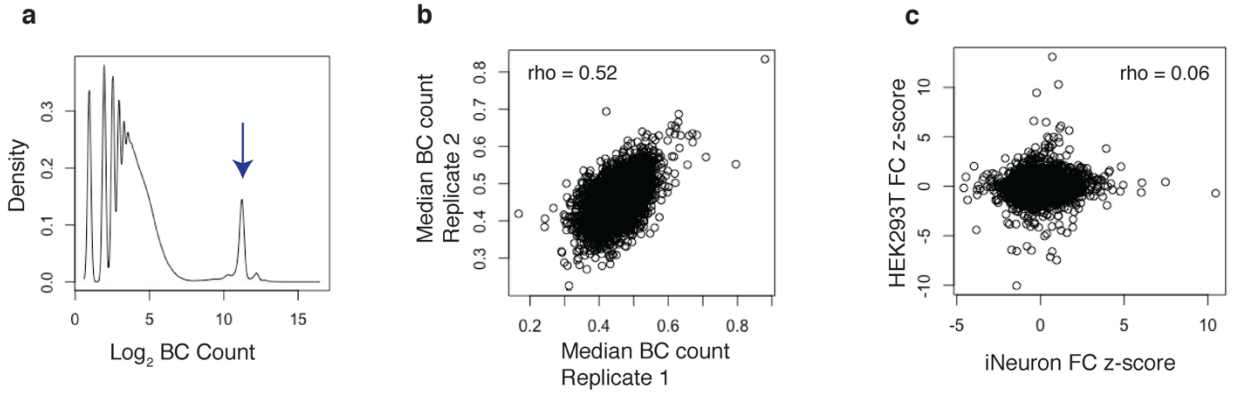


Figure 6-5. AAV-MPRA exhibited poor technical performance in iNeurons. (**A**) Density plot of log_2 normalized barcode (BC) counts shows a peak (blue arrow) at extremely high counts, indicative of PCR overamplification. (**B**) Representative plot displaying low allele level correlations between replicates (Spearman's rho). (**C**) There was little correlation between variant effect size z-scores (e.g. FC z-scores) when the library was tested in HEK cells vs. IPSC-derived neurons (iNeurons).

NUCLEOFECTION PROTOCOL

The following is a summary of work performed in collaboration with Dr. Qiuyu Guo in the Geschwind lab. We opted to test a nucleofection protocol in hNPCs as an alternative to AAV delivery. First, we noted previous reports that plasmid transfection in primary cell lines can activate the cGAS-STING pathway and drive the type-I interferon response [33]. Type-I interferons can induce a highly coordinated pro-inflammatory transcriptional response, potentially biasing our MPRA results [33]. We first confirmed that plasmid nucleofection in hNPCs induced robust expression of IFN-response genes and found that this inflammatory response could be inhibited

by co-incubation with the inhibitors C16 and BX-795 (Figure 6-6a) [33]. Next, we tested whether our MPRA libraries could be nucleofected into hNPCs. We nucleofected 20 ug of library into six biological replicates of 5 million hNPCs each, in the presence of C16 and BX-795 inhibitors. Three replicates were collected 24 hours later (undifferentiated), while another three replicates were differentiated into post-mitotic neurons for 14 days before collection. Then, we isolated mRNA and plasmid DNA before performing deep sequencing. Library quality was assessed by looking at inter-replicate variant level correlation. Unfortunately, the differentiated hNPC libraries exhibited very poor quality, with mean correlation of 0.2. However, the undifferentiated library had modest inter-replicate correlations of 0.77-0.82 (Pearson's R^2; Figure 6-6b). While these results remain sub-optimal, they represent a dramatic improvement over the AAV-delivery method and suggests that nucleofection could be a promising approach. Further optimization of this protocol remains ongoing.

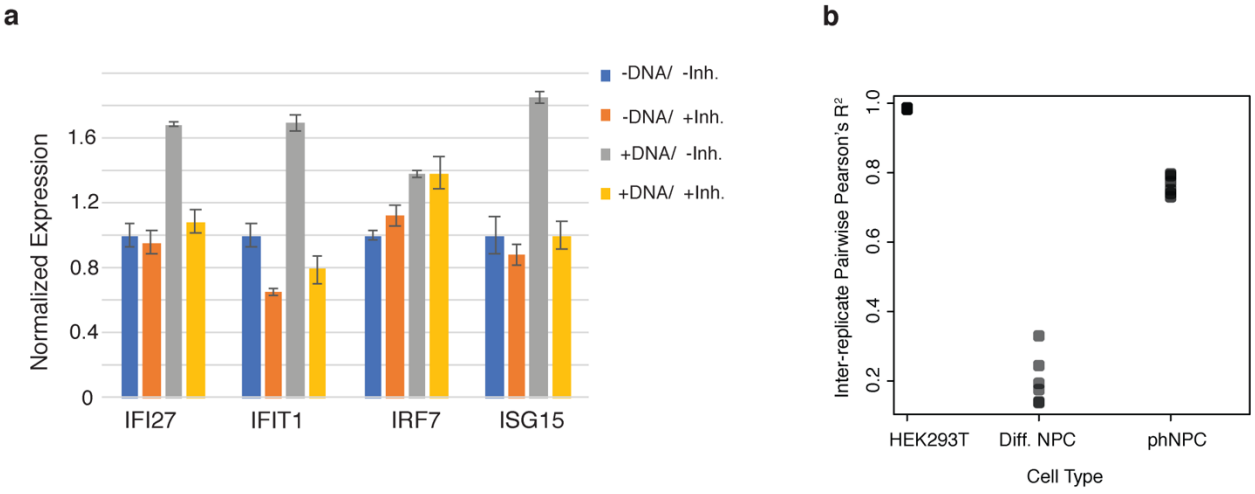


Figure 6-6. Nucleofection of MPRA libraries into neural progenitor cells (NPCs). (**A**) NPCs were placed into four different treatment groups: +/- plasmid nucleofection and +/- 0.1 uM of C16 and BX-795 inhibitors. Interferon activity was then assessed through quantitative PCR of four downstream genes: IFI27, IFIT1, IRF7, and ISG15 which were normalized against GAPDH expression. Plasmid DNA nucleofection significantly increased IFN activity (grey bars), which could be rescued by incubation with the inhibitors (yellow). (**B**) MPRA library performance (assessed as allele-level inter-replicate Pearson's correlations) across three different cell types: HEK293T, differentiated hNPCs, and undifferentiated hNPCs.

Discussion

The recent development and deployment of Massively Parallel Reporter Assays has enabled the widespread functional characterization of noncoding genetic elements. However, the contribution of a number of technical parameters on assay performance has remained relatively undefined. In this work, I use my MPRA data to both reproduce previous findings and provide novel insights into factors germane to MPRA implementation (Table S6-1 summarizes these findings). First, as was noted in Klein *et al.* [9], I demonstrate that the "classic" MPRA design based on the pGL4 reporter construct [2] is highly reproducible and robust to technical noise across separate experiments. I also find that detection power for allelic effects is especially dependent on a high degree of barcode complexity. In agreement with Tewhey and colleagues [5], I show that approximately 40 barcodes per allele maximizes sensitivity, though this may be dependent on specific analysis methods.

Additionally, I describe technical factors that influence waste and overall cost. I find that sequence features including GC content extremes, uninterrupted strings of dinucleotide repeats, and low sequence complexity predict element dropout during library construction. Prefiltering elements based on extremes of these predictor values is warranted, freeing up limited space within synthesis microarrays that might otherwise be wasted. I also note that a relatively low sequencing coverage of 20-25 fold is sufficient to ensure assay quality. However, I find that a large percentage (~50%) of reads are wasted due to mapping to "junk" library elements introduced during element synthesis and cloning. Remediation of this issue will largely depend on further fidelity improvements for DNA synthesis.

By assessing variants in multiple element configurations, including differing amounts of upstream and downstream genomic context, I show that distal DNA sequences have a modest

impact on allelic effects (r = 0.78). Likewise, Klein *et al.* [9] note that increasing enhancer size changes regulatory function through the introduction of distal TFBSs with transcriptional modifying effects. Currently, DNA elements are typically obtained by microarray synthesis, which limits element size to around 200 bp [9]. However, this likely underestimates the true width and relevant regulatory context of most enhancers. Alternative methods for DNA capture and improved oligo assembly methods may rectify this. Additionally, Klein and colleagues report minimal effects of element orientation on performance (mean r = 0.88) [9], while I find a stronger effect from this parameter (mean r = 0.69). However, they primarily assessed enhancer elements, while noting that element asymmetry from promoters is stronger. It is possible that orientation effects are minimal for enhancer elements but more influential in elements derived from other genomic features such as promoters.

Finally, by comparing four different MPRA libraries requiring differing numbers of PCR amplification cycles prior to sequencing (a proxy for library mRNA abundance), I find that robust library coverage is the single most critical factor contributing to assay quality. This observation underlies an inherent tension in the technique, which is that reporter assays are highly dependent on cellular contexts but problematic to implement across most primary cell types, due to the difficulty of large scale cell cultures and efficient library transduction. Overall library expression can be roughly captured by the equation: Expression = Total Cell Number × Transduction Percentage × Copy Number × Transcriptional Drive. And: Coverage = Expression / Library Complexity. Transcriptional drive is an innate feature of the cell type and particular library used. Therefore, coverage (and performance by extension) can be improved by increasing cell numbers per replicate, maximizing delivery efficiency, or reducing library complexity (e.g.

fewer elements tested). I surveyed variants associated with neurodegenerative disease, which ideally would be tested within human neuronal model systems.

In Part 2 I discussed my attempts to implement MPRA within these cell types, which are notoriously difficult to transduce. The first approach, which entailed packaging my MPRA library into AAVs for subsequent delivery into IPSC-derived neurons, failed to generate usable quality data. Although I screened multiple AAV serotypes identifying PHP.eB [27] as having the best neuronal tropism, this synthetic capsid was originally developed using directed evolution in the murine nervous system. Similarly, other serotypes with reported neuronal tropism are almost uniformly test in rodents [42]. It is likely that these AAV serotypes are not optimized for transduction of human neurons, and the field would greatly benefit from the development of novel capsids for this application. We also tested a nucleofection delivery strategy in hNPCs. For differentiated neurons, data quality was again quite poor. Nucleofected plasmids remain episomal, so residual post-transfection cell division will effectively "dilute" the library. Additionally, library performance could worsen through epigenetic silencing. By contrast, data quality from the undifferentiated condition was relatively promising. We intend to further improve quality by doubling or tripling cell numbers per condition to boost library coverage and quality. Finally, an unexplored approach thus far is to use lentivirus for library delivery, which has the advantage of integrating into the genome [9,12]. In contrast to transfection or AAV infection which can achieve high copy number (10-100s), lentivirus achieves a more modest number of integrations (1-10) per cell. Thus, this approach would require very large cell numbers to achieve adequate coverage for measuring allelic effects, but could nevertheless be viable for relatively small libraries of variants.

Supplement

Technical Parameter	Significance	Source	Additional Notes
Oligo sequence features	+	This work	Extreme GC content, low sequence complexity, and nucleotide repeats predict dropout. Can drop such oligos before array synthesis.
Enhancer placement: 5' vs. 3'	++/+++	Klein et al., 2020 [9]	5' placement is more sensitive to promoter binding factors. 3' placement is sensitive to RNA binding proteins.
Distal genomic context of element	++	Klein et al., 2020 [9] This work	Klein et al. assessed this by looking at enhancers of different lengths. Currently limited by microarray synthesis.
Element orientation	+ or ++	Klein et al., 2020 This work	Contradictory results: may depend on the type of elements tested.
Choice of minimal promoter	Variable	Ernst et al., 2016 [16] Jayavelou et al., 2019 [18]	Depends on experimental goals. Weak minimal promoter is good for detecting active regions, strong minimal promoter is good for detecting repressive regions.
Integrating vs. episomal assay	++	Inoue et al., 2017 [12] Klein et al., 2020 [9]	
Barcode complexity	++/+++	Tewhey et al., 2016 [5] This work	Especially important for power to detect variants with small effect sizes. Less relevant to enhancer screens.
Number of biological replicates	++/+++	Myint et al., 2019 [15]	Also relevant for power, good saturation achieved at 6 replicates.
Library Coverage / Expression	++++	Melnikov et al., 2014[19] This work	Cellular transduction and expression of the MPRA library. This is the single most critical parameter.
Assay cell type	+++	Mulvey et al., 2020 [1]	Cell type specific *trans* factors greatly influence results, rather than data quality.
Sequencing depth	+	This work	Quality maintained at relatively low coverage
Analysis pipeline	Variable	Myint et al., 2019 [15]	Depends on application. There are also clearly incorrect methods to avoid, such as Chi-square tests.

Table S6-1. Summary of technical parameters influencing MPRA performance. Each parameter is scored on its relative impact: + Minor, ++ Moderate, +++ Strong, ++++ Critical. Sources informing these findings as well as additional notes are also listed.

Bibliography

1. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts. *Biol. Psychiatry* (2020).
2. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265 (2012).
3. 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).
4. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
5. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
6. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).
7. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 1–14 (2017).
8. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
9. 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).
10. Trauernicht, M., Martinez-Ara, M. & van Steensel, B. Deciphering Gene Regulation Using Massively Parallel Reporter Assays. (2019).

11. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

12. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

13. Davis, J. E. *et al.* Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. *Cell Syst.* **11**, 75–85 (2020).

14. Ghazi, A. R. *et al.* Design tools for MPRA experiments. *Bioinformatics* **34**, 2682–2683 (2018).

15. Myint, L., Avramopoulos, D. G., Goff, L. A. & Hansen, K. D. Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* **20**, 1–19 (2019).

16. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).

17. Castaldi, P. J. *et al.* Identification of functional variants in the FAM13A chronic obstructive pulmonary disease genome-wide association study locus by massively parallel reporter assays. *Am. J. Respir. Crit. Care Med.* **199**, 52–61 (2019).

18. Jayavelu, N. D., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1–15 (2020).

19. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. Massively parallel reporter assays in cultured mammalian cells. *JoVE J. Vis. Exp.* e51719 (2014).

20. Kalita, C. A. *et al.* QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* **34**, 787–794 (2018).

21. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 1–17 (2019).

22. Qiao, D. *et al.* Statistical considerations for the analysis of massively parallel reporter assays data. *Genet. Epidemiol.* **44**, 785–794 (2020).

23. Maricque, B. B., Dougherty, J. D. & Cohen, B. A. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res.* **45**, e16–e16 (2017).

24. Caballero, J., Smit, A. F., Hood, L. & Glusman, G. Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic Acids Res.* **42**, e99–e99 (2014).

25. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

26. Dalkara, D. *et al.* In vivo–directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. *Sci. Transl. Med.* **5**, 189ra76-189ra76 (2013).

27. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).

28. Sørensen, A. T. *et al.* A robust activity marking system for exploring active neuronal ensembles. *Elife* **5**, e13918 (2016).

29. Stein, J. L. *et al.* A quantitative framework to evaluate modeling of cortical development by neural stem cells. *Neuron* **83**, 69–86 (2014).

30. Wang, C. *et al.* Scalable production of iPSC-derived human neurons to identify tau-lowering compounds by high-content screening. *Stem Cell Rep.* **9**, 1221–1233 (2017).

31. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 1–19 (2016).

32. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).

33. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141 (2018).

34. Earley, L. F. *et al.* Adeno-associated virus serotype-specific inverted terminal repeat sequence role in vector transgene expression. *Hum. Gene Ther.* **31**, 151–162 (2020).

35. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).

36. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, 1–17 (2014).

37. Consortium, Gte. Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).

38. Maloney, B. & Lahiri, D. K. Structural and functional characterization of H2 haplotype MAPT promoter: unique neurospecific domains and a hypoxia-inducible element would enhance rationally targeted tauopathy research for Alzheimer's disease. *Gene* **501**, 63–78 (2012).

39. Swarup, V. *et al.* Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nat. Med.* **25**, 152–164 (2019).

40. Swarup, V. *et al.* Identification of conserved proteomic networks in neurodegenerative dementia. *Cell Rep.* **31**, 107807 (2020).

41. Zincarelli, C., Soltys, S., Rengo, G. & Rabinowitz, J. E. Analysis of AAV serotypes 1–9 mediated gene expression and tropism in mice after systemic injection. *Mol. Ther.* **16**, 1073–1080 (2008).

42. Duong, T. T. *et al.* Comparative AAV-EGFP transgene expression using vector serotypes 1–9, 7M8, and 8b in human pluripotent stem cells, RPEs, and human and rat cortical neurons. *Stem Cells Int.* **2019**, (2019).

CHAPTER 7

Conclusions and future directions

Conclusions

In this work, I perform a comprehensive characterization of common genetic variation associated with two neurodegenerative disorders that share overlapping clinicopathologic features – Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP). I use Massively Parallel Reporter Assays - a high-throughput experimental approach [1,2] - to screen 5,706 variants derived from three genome wide association studies for these disorders [3–5], identifying 320 regulatory polymorphisms distributed across 27 of 34 tested loci. Confidence in these findings is supported by confirming that: 1) the assay is highly reproducible and robust to technical noise, 2) MPRA-defined regulatory variants are enriched within functional genomic annotations derived from human brain, and 3) four of six (66%) regulatory predictions were verified using genome editing within brain-relevant cell lines. Thus, I conclude that Massively Parallel Reporter Assays efficiently survey and prioritize functional regulatory variants within GWAS loci for brain related traits.

This work demonstrates the utility of identifying underlying regulatory variants within GWAS loci [6]. Isolating causal variants enables downstream prediction of genes and functional mechanisms connecting phenotypes with associated risk loci. In particular, complex genomic regions harboring extensive linkage disequilibrium and multiple genes substantially benefit from such functional analyses, as these loci are particularly ambiguous and difficult to characterize using traditional statistical approaches. Here I highlight three such regions, the pan-neurodegeneration risk locus in 17q21.31 (containing MAPT and primarily associated with PSP) [7], and the 19q13.32 (APOE) and 6p21 (HLA) regions associated with AD [4]. I identify regulatory variants distributed across these loci that regulate novel putative risk genes, including *PLEKHM1* and *C4* within the 17q21.31 and 6p21 regions, respectively.

Furthermore, identifying underlying causal variants allows for the more precise characterization of genetic risk factors shared across trait-associated loci. As discussed in Chapter 1, GWAS loci are mostly composed of neutral trait-associated variants due to LD [6,8]. While many studies have identified enrichment of GWAS variation within functional genomic features to uncover trait-relevant biology [9–13], the presence of many non-functional variants undoubtedly introduces noise into such analyses. Here, I demonstrate that MPRA-defined regulatory variants preferentially disrupt binding sites for transcription factors that form cell-type and disease-specific regulatory networks, *relative to the complement set of non-functional GWAS variants*. In particular, this implicates dysregulation of a neuronal transcriptional network composed of SP1 and its binding partners as a genetic risk factor for PSP (Chapter 4). Although these findings remain to be verified, such analyses are made possible by specifically considering refined sets of regulatory variants.

Limitations

Of note, identification of transcriptional regulatory variants by MPRA does not directly imply causality. In the context of GWAS interpretation, loci may contain multiple variants with distinct molecular functions of which only a subset meaningfully influence the phenotype of interest. For example, it has been estimated that there is a roughly proportional phenotypic impact between common variation affecting mRNA splicing (sQTL) and gene expression (eQTL) [14]. The identification of both eQTLs and sQTLs within a given GWAS locus does not entail that both mechanisms are causal, nor does the presence of multiple differentially regulated genes within a risk haplotype entail that all genes are causal. Thus, the causal determination of regulatory predictions should be validated through careful genetic modeling using *in vitro* assays

or animal models. Moreover, MPRAs to screen allelic effects such as the one utilized in this work cannot capture variation influencing other genetic mechanisms (chromatin accessibility, splicing, RNA stability, etc.), which must be identified through other means.

A limitation specific to this work was the performance of MPRA within HEK293T cells. The *trans*-cellular environment substantially contributes to reporter assay outcomes [15–17] and undoubtedly meaningfully differs between HEK293T and primary brain cell types. I addressed this limitation by integrating brain specific functional annotations with my MPRA results and by performing genome editing in relevant brain cell types to validate four of six tested predictions. I also found a 60% overlap between GWAS variation residing within open chromatin in brain cell types and HEK293T cells (Chapter 2). Taken together, these results suggest that my MPRA is valuable as a high-throughput preliminary screen, but has the potential to miss a proportion of tested variants. Nevertheless, as discussed in Chapter 3, MPRA provides valuable information orthogonal to other approaches for variant prioritization. Thus, while MPRA is not a panacea, it is an important component of a broader arsenal of complementary methods for the regulatory annotation and functional interpretation of noncoding loci and variants [6,8].

Future directions

These limitations underscore the need for future work to solidify and expand upon my findings. First, while I was able to verify select predictions from my screen using genome editing, a number of important loci remain to validated. This process could be parallelized by performing a CRISPRi screen, ideally in IPSC-derived neurons or organoids [18], to transcriptionally inhibit regions containing predicted regulatory variants and verify suppression of cognate gene expression. True gold standard validation of variant function would ideally be

performed using CRISPR/Cas9 mediated allelic replacement, though these are admittedly time-consuming and laborious experiments that are at the cutting edge of feasibility . Second, this work implicates a number of novel candidate risk genes for PSP and AD (e.g. *PLEKHM1, APOC1, C4*). The relationship between these genes and disease pathogenesis should be clarified through careful mechanistic studies in relevant disease models.

Furthermore, an ongoing area of research is the implementation of MPRAs screening variation associated with brain-related traits within neuronal model systems that more closely recapitulate the relevant cellular environment. This would improve the sensitivity, specificity, and interpretability of these screens. More broadly, MPRA tests genomic elements artificially removed from their natural chromatin environment, and it cannot simultaneously screen multiple regulatory mechanisms. In the future, the field should move towards high-throughput screens that directly perform allelic replacement in the mammalian genome. This has been done on a limited basis in pioneering studies performing saturation mutagenesis, [19,20] or multiplexed base editing [21]. However, because the efficiency of allelic editing is low, these approaches require phenotypic selection to enrich for functional mutations, thereby limiting experimental flexibility. Future technical advances improving the efficiency of allelic replacement or the single cell identification of genetic mutations would facilitate the development of next generation approaches that parallelized annotation of noncoding variation assessed directly within the mammalian genome.

Bibliography

1. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265 (2012).

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. Chen, J. A. *et al.* Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

4. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452 (2013).

5. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

6. Cano-Gamez, E. & Trynka, G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* **11**, (2020).

7. Bowles, K. *et al.* 17q21. 31 sub-haplotypes underlying H1-associated risk for Parkinson's disease and progressive supranuclear palsy converge on altered glial regulation. (2019).

8. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

9. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

10. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

11. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, (2014).

12. de la Torre-Ubieta, L. *et al.* The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* **172**, 289–304 (2018).

13. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

14. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

15. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).

16. Maricque, B. B., Dougherty, J. D. & Cohen, B. A. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res.* **45**, e16–e16 (2017).

17. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts. *Biol. Psychiatry* (2020).

18. Kampmann, M. CRISPR-based functional genomics for neurological disease. *Nat. Rev. Neurol.* **16**, 465–480 (2020).

19. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

20. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).

21. Hanna, R. E. *et al.* Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064–1080 (2021).

CHAPTER 8: Appendix A

Supplemental Materials and Methods

Table A-1. Primers

MPRA library construction	**Sequence (5' - 3')**
barcode_new_F	/5Biosg/CTGAGTACTGTATGGGCGA
barcode_N_R	/5Biosg/AGTCGACTAGTNNNNNNNNNNNNNNNNNNNTCTAGAACAGTTGA CACTTTTGTCGG
BC_map_P5_Rev	AATGATACGGCGACCACCGAGATCTACACGTAACCACCCTGATCGACGG
BCmap_P7_For	CAAGCAGAAGACGGCATACGAGATTCGGCAGTTGGGAAGAGCATAGTCG
BCmap_R1Seq_Rev	GTAACCACCCTGATCGACGGGGAGTGTACTAGT
BCmap_R2Seq_For	TCGGCAGTTGGGAAGAGCATAGTCGTAGAGCACGCGT
Amp_minPLuc2_For	/Biosg/ACGACGTTGTAAAACGACGG
Amp_minPLuc2_Rev	/Biosg/CACAGGAAACAGCTATGACC
Lib_Hand_RT	ATGCTCTTCCCAACTGCCGACGACGGGGAGTGTACTAGT
Lib_Hand	ATGCTCTTCCCAACTGCCGA
Lib_seq_eGFP_F2	CAAGATCCGCCACAACATCG
P5_seq_eGFP_F2	AATGATACGGCGACCACCGAGATCTACACCAAGATCCGCCACAACATCG
P7_Ind_#_Han	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGCGTGCTCTACGACTATGC TCTTCCCAACTGCCGA
Exp_eGFP_Seq_F2	GGATCACTCTCGGCATGGACGAGCTGTACAAGTAATCTAGA
Exp_Ind_Seq_P	TCGGCAGTTGGGAAGAGCATAGTCGTAGAGCACGC
Genotyping	
rs636317_genomic_F	TTTGGCTGTGAATCCGTCTGG
rs636317_genomic_R	TAGAACAAAGCCCGACACAGTG
rs13025717_genomic_F	AAGTCTGCAGAAAGGTAGC
rs13025717_genomic_R	GCATGCCCATAGAACTTGG
rs6064392_genomic_F	TCTTGCCCAGCCAATTCATG
rs6064392_genomic_R	GAAGAGCCTGCCCTTAAGC
rs7920721_genomic_F	CCTGTAATCCCAGCTACTC
rs7920721_genomic_R	AGAGATCGAGACCATCCTG

rs9271171_genomic_F	AGAGGAGTACTGTCTGATGGG
rs9271171_genomic_R	CCAGAATTCTGACCTCATGACC
rs1532277_genomic_F	CATTAAGCACAGGCTAGACC
rs1532277_genomic_R	CCAATCAGGGAAGTAAGTACG
qPCR	
BIN1_F	ACAACGACCTGCTGTGGATGG
BIN1_R	CGTGACTTGATGTCGGGGAACT
MS4A6_F	TGTTCCCAATGAGACCATCA
MS4A6_R	AGCAGATGCCAAAATGATCC
GAPDH_F	TCGACAGTCAGCCGCATCTTCTT
GAPDH_R	GCGCCCAATACGACCAAATCC
ACTB_F	CACCATTGGCAATGAGCGGTTC
ACTB_R	AGGTCTTTGCGGATGTCCACGT
C4A_F	CCTGAGAAACTGCAGGAGACAT
C4A_R	GTGAGTGCCACAGTCTCATCAT
HLA-DRB1_F	CACCAGACCACGTTTCTTGGAGT
HLA-DRB1_R	CACGTTCTCCTCCTGGTTATGGA
HLA-DQA1_F	CTTGCCCTGACCACCGTGATGA
HLA-DQA1_R	CAGAGGGACCGTAAGACTGGTAC
USP6NL_F	ATACTCAGCCTTTCAACTCG
USP6NL_R	GCAAGTACACGTCAAATCTC
ECHDC3_F	CAAGTCCTCTTTTGCCACTCC
ECHDC3_R	ATCTCCAAGGCCACCTTTCTA
CASS4_F	GGGTTGGTGGAAGTGTTTGC
CASS4_R	TCTTCCAGGCCTCTCAGGAA
RTF2_F	TGCTGAAGACAAGGATGGAG
RTF2_R	TGAAACAGACTCTGCTGCCT
sCLU_F	AGGCGTGCAAAGACTCCA
sCLU_R	GCCCACTCTCCCAGGTCA

Table A-2. CRISPR gRNAs (Chapter 5)

SNP	ID	5' guide	3' guide	sequence (5' - 3')
rs636317	g15	X		ATGACACAGAGTCATGCCAA
rs636317	g16		X	GTACCCGAAAATCACTGGAG

rs13025717	g11	X		AGACTGAGTTGGAAAACGGA
rs13025717	g12		X	AACAGGACCTCACTGTCACG
rs13025717	g14		X	ACGCACCATGCTTAGCAACA
rs6064392	g25	X		TGCCTTATACATGCGTAGGG
rs6064392	g26		X	TATTCACAAAATATCAAACT
rs7920721	g21	X		GAAACCGCAGCCCATGAGCC
rs7920721	g20		X	TGCACTCTGTCCTGGCGACA
rs9271171	g3	X		TCCTAGACTTGTAACTACAC
rs9271171	g4		X	TAGTTTATTTGAGATCAGCA
rs9271171	g5	X		AATATTCTCATAATCATGCT
rs9271171	g6		X	ATTGTCCTATGACAATCAGC
rs1532277	g7	X		CAGAACTCTAGCAAGACGTG
rs1532277	g8		X	CCAGTGGGATGGTCAAGGCA
rs1532277	g9	X		TCAGGAAGCTTATCTAATAG
rs1532277	g10		X	ATTTGCTTCTGAAAGCATCA

Table A-3. ENCODE Accessions [1,2] (Chapter 2)

Sample Description	**Accession**
HEK293T DNase-seq hotspot	ENCFF013WVF
bipolar neuron DNase-seq hotspot rep 1	ENCFF017HNT
bipolar neuron DNase-seq hotspot rep 2	ENCFF502TUB
brain pericyte DNase-seq hotspot	ENCFF133GOH
frontal cortex DNase-seq hotspot rep 1	ENCFF661TYD
frontal cortex DNase-seq hotspot rep 2	ENCFF861YPP
neural progenitor DNase-seq hotspot rep 1	ENCFF800OYT
neural progenitor DNase-seq hotspot rep 2	ENCFF077TNH
astrocyte DNase-seq hotspot rep 1	ENCFF606JSS
astrocyte DNase-seq hotspot rep 2	ENCFF529ZNC
CD14+ monocyte DNase-seq hotspot rep 1	ENCFF281ASX
CD14+ monocyte DNase-seq hotspot rep 2	ENCFF943FMD
HEK293T DHS	ENCFF910QHN
HEK293 H3K27ac replicated narrowPeak	ENCFF668WID
HEK293 H3K4me3 replicated narrowPeak	ENCFF728WLM
HEK293 H3K36me3 pseudo-replicated narrowPeak	ENCFF496OIF

HEK293 H3K9me3 replicated narrowPeak	ENCFF037SXA
HEK293T SP1 optimal IDR narrowPeak	ENCFF240PYU
HEK293T SP2 optimal IDR narrowPeak	ENCFF905HYT
HEK293T ARNT optimal IDR narrowPeak	ENCFF550UEU
HEK293 KLF14 optimal IDR narrowPeak	ENCSR780ESQ
HEK293T DNase-seq narrowPeak	ENCFF910QHN
astrocyte DNase-seq narrowPeak rep 2	ENCFF803JHO
astrocyte DNase-seq narrowPeak rep 2	ENCFF399UZY
neural progenitor DNase-seq narrowPeak rep 1	ENCFF572AAL
neural progenitor DNase-seq narrowPeak rep 2	ENCFF230SPN
bipolar neuron DNase-seq narrowPeak rep 1	ENCFF524XCQ
bipolar neuron DNase-seq narrowPeak rep 2	ENCFF950DMO
fetal brain DNase-seq narrowPeak rep 1	ENCFF528GDM
fetal brain DNase-seq narrowPeak rep 2	ENCFF457XYZ
CD14+ monocyte DNase-seq narrowPeak rep 1	ENCFF063IUG
CD14+ monocyte DNase-seq narrowPeak rep 2	ENCFF815GDP
CD14+ monocyte H3K27ac BigWig	ENCFF437JSB
CD14+ monocyte H3K4me3 BigWig	ENCFF485XYG
CD14+ monocyte H3K4me1 BigWig	ENCFF929UOK
CD14+ monocyte CTCF BigWig	ENCFF971FFO
CD14+ monocyte DNase-seq narrowPeak	ENCFF376RLJ

Table A-4. External Data Access Links

Description	Access/Download Link
External MPRA Data (Tewhey et al., 2016) [3]	https://www.cell.com/cms/10.1016/j.cell.2016.04.027/attachment/ddbd23af-33df-41ff-a796-5b89759b0a97/mmc2.xlsx
CADD [4]	https://cadd.gs.washington.edu/score
GWAVA [5]	https://www.sanger.ac.uk/sanger/StatGen_Gwava
CATO [6]	http://www.mauranolab.org/CATO/
LINSIGHT [7]	http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw
SNPS2TFBS [8]	http://ccg.vital-it.ch/snp2tfbs/
Microglia H3K27ac [9]	http://homer.ucsd.edu/hubs//nuclei_h3k27ac_hg19_pooled/hg19/human_PU1nuclei_H3K27ac_epilepsy_pooled_hg19.ucsc.bigWig
Neuron H3K27ac	http://homer.ucsd.edu/hubs//nuclei_h3k27ac_hg19_pooled/hg19/human_NEUNnuclei_H3K27ac_epilepsy_pooled_hg19.ucsc.bigWig
Oligodendrocyte H3K27ac	http://homer.ucsd.edu/hubs//nuclei_h3k27ac_hg19_pooled/hg19/human_OLIG2nuclei_H3K27ac_epilepsy_pooled_hg19.ucsc.bigWig
Astrocyte H3K27ac	http://homer.ucsd.edu/hubs//nuclei_h3k27ac_hg19_pooled/hg19/human_LHX2nuclei_H3K27ac_epilepsy_pooled_hg19.ucsc.bigWig
Microglia H3K4me3	http://homer.ucsd.edu/hubs//nuclei_h3k4me3_hg19_pooled/hg19/human_PU1nuclei_H3K4me3_epilepsy_hg19.ucsc.bigWig

Neuron H3K4me3	http://homer.ucsd.edu/hubs//nuclei_h3k4me3_hg19_pooled/hg19/human_NEUNnuclei_H3K4me3_epilepsy_hg19.ucsc.bigWig
Oligodendrocyte H3K4me3	http://homer.ucsd.edu/hubs//nuclei_h3k4me3_hg19_pooled/hg19/human_OLIG2nuclei_H3K4me3_epilepsy_hg19.ucsc.bigWig
Astrocyte H3K4me3	http://homer.ucsd.edu/hubs//nuclei_h3k4me3_hg19_pooled/hg19/human_LHX2nuclei_H3K4me3_epilepsy_pooled_hg19.ucsc.bigWig
Microglia enhancers	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/PU1_enhancers.sorted.bigWig
Neuron enhancers	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/NeuN_enhancers.sorted.bigWig
Oligodendrocyte enhancers	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/Olig2_enhancers.sorted.bigWig
Astrocyte enhancers	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/LHX2_enhancers.sorted.bigWig
Microglia promoters	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/PU1_promoters.sorted.bigWig
Neuron promoters	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/NeuN_promoters.sorted.bigWig
Oligodendrocyte promoters	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/Olig2_promoters.sorted.bigWig
Astrocyte promoters	http://homer.ucsd.edu/iholtman/Nuclei_project/peak_files_IDR_hg19/UCSC_peak_files/hg19/LHX2_promoters.sorted.bigWig
HACER cell type enhancers [10]	http://bioinfo.vanderbilt.edu/AE/HACER/download/T1.txt
THP-1 Hi-C loops [11]	http://promoter.bx.psu.edu/hi-c/downloads/loops-hg19.zip
Gene annotations	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/genes/hg19.ensGene.gtf.gz
RNA HPA cell line gene data [12]	https://www.proteinatlas.org/download/rna_celline.tsv.zip
Single Cell Human Brain RNA-seq from Allen Brain Atlas	https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x
List of human TFs	http://humantfs.ccbr.utoronto.ca/download/v_1.01/TF_names_v_1.01.txt

Bibliography

1. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
2. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).
3. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

4. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

5. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of non-coding sequence variants. *Nat. Methods* **11**, 294–296 (2014).

6. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393 (2015).

7. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).

8. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS–a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).

9. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).

10. Wang, J. *et al.* HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).

11. Phanstiel, D. H. *et al.* Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. *Mol. Cell* **67**, 1037–1048 (2017).

12. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).