Genome-wide landscape of RNA-binding protein dysregulation reveals a major impact on psychiatric disorder risk

Christopher Y. Park^{1,*}, Jian Zhou¹, Aaron K. Wong¹, Kathleen M. Chen¹, Chandra L. Theesfeld², Robert B. Darnell^{3,4}, Olga G. Troyanskaya^{1,2,5,*}

¹Flatiron Institute, Simons Foundation, New York, New York, United States of America

²Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America

³Laboratory of Molecular Neuro-Oncology, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.

⁴Howard Hughes Medical Institute

⁵Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America

Correspondence to:

Olga G. Troyanskaya, email: ogt@cs.princeton.edu

Christopher Y. Park, email: cpark@flatironinstitute.org

Abstract

Despite the strong genetic basis of psychiatric disorders, the molecular origins of these diseases are still largely unmapped. RNA-binding proteins (RBPs) are responsible for most post-transcriptional regulation, from splicing to translational to localization. RBPs thus act as key gatekeepers of cellular homeostasis, especially in the brain. Here, we leverage a deep learning approach to interrogate variant effects genome-wide, and discover that the dysregulation of RBP target sites is a principal contributor to psychiatric disorder risk. We show that specific modes of RBP regulation are genetically linked to the heritability of psychiatric disorders, and demonstrate that diverse RBP regulatory functions are reflected in distinct genome-wide negative selection signatures. Notably, RBP dysregulation has a stronger impact on psychiatric disorders than common coding region variants and explains heritability not currently captured by large-scale molecular QTL studies (expression QTLs and splicing QTLs). We share genome-wide profiles of RBP target site dysregulation, which we used to identify DDHD2 as a candidate schizophrenia risk gene, in a public web server. This resource provides a novel analytical framework to connect the full range of RNA regulation to complex disease.

Main

Interrogating the genetics behind psychiatric disorders is a key path in understanding the pathophysiological cause of mental illness, in part due to the largely immutable nature of the genome during the pathogenic manifestation. In particular, genome-wide associated studies (GWAS) have been the most widely adopted human genetics approach for studying psychiatric disorders. With ever-increasing sample cohorts, surpassing one million subjects for some psychiatric traits¹, numerous risk loci have now been cataloged. Despite this progress, the pathological mechanisms and biochemical perturbations that lead to human psychiatric disorders are still lacking – a critical step for translating the genetic discoveries into actionable targets.

RNA-binding proteins (RBP) regulate various aspects of RNA metabolism, including RNA splicing², localization³, stability⁴ and translation⁵. Each of these functions are critical not only in the expression of proteins, but in establishing their proper spatiotemporal function, especially in the brain⁶. Proper regulation of the full RNA life cycle by RBPs is particularly important in neurobiology, where complex regulatory events take place in synapses far away from the nucleus⁷.

RBP-encoding genes are frequently mutated in neuropsychiatric disorders, suggesting a pathogenic role⁸⁻¹⁰ and inspiring follow-up efforts to study the larger set of variants at the RNA targets that affect RBP-RNA interactions¹¹⁻¹⁵. These studies have shed light on the roles of specific RBPs and functions, particularly splicing, in neurological disease. However, comprehensive and genome-wide insight into diverse RBPs, functions and causal target sites are

lacking. Thus, to what extent dysregulated RNA-RBP interactions contribute to psychiatric disorders remains an open question.

Here, we address this challenge with the first genome-wide, systematic analysis of the role of RBP target site dysregulation in psychiatric disease. To map variant impacts on RBP-RNA interactions at scale, we leverage a deep learning-based sequence model, Seqweaver, whose accuracy in predicting variant effects on RBP target site dysregulation we previously extensively evaluated both computationally and experimentally, including by detecting *de novo* noncoding mutation signal in autism in probands versus their unaffected siblings¹⁶. We apply Seqweaver to profile allele-specific effects of inherited variants genome-wide, enabling us, for the first time, to examine the diverse landscape and impact of RBP dysregulation in complex psychiatric disorders.

Our study generates genome-wide variant level annotations of RBP dysregulation, which we make publicly available at hb.flatironinstitute.org/seqweaver. We show that the dysregulation of RBP target sites and diverse RBP regulatory functions are top drivers of psychiatric disorder risk. This analytical framework will accelerate biochemical investigation into disease genetics. Indeed, leveraging this resource as a case study, we discover a novel link between an RBP-disrupting variant in DDHD2, which encodes a phospholipase involved in hereditary spastic paraplegia, and a multiethnic-associated locus that increases the risk of schizophrenia.

Figure 1. Negative selection signatures differentiate RBPs by their regulatory function.

a) For each RBP, the negative selection signatures along the major axes of sub-genic regulatory regions (5'/3' UTR and introns 200 bp flanking alternatively spliced exon) are plotted. Higher coefficient z-scores, along the x, y or z axes, implies stronger regional selection compared to the gene background. The fitness effects contributed by RBP regulatory roles, beyond splicing, are shown by the separation of RBPs along the regional selection axes. b) Examples of regional sub-genic selection signatures for the three major noncoding regions regulated by RBPs. GnomAD cohort noncoding variants (MAF bins x-axis) and variant set mean RBP dysregulation estimates (Y-axis) are shown. Stronger RBP selection signatures can be observed by the larger slope for each sub-genic region compared to the background gene levels. All inferred mean RBP dysregulation scores were normalized by subtracting average dysregulation predicted scores of common variants (>0.05 MAF) for comparison (95% CI).

Results

Regulatory function of RBPs is reflected in their genome-wide negative selection signature

Mutations that cause neuropsychiatric disorders are expected to decrease fitness, with epidemiology studies showing that these patients have reduced fecundity compared to the general population¹⁷. Consequently, we hypothesized that if RBP dysregulation contributes to brain-associated diseases, then variants disrupting RBP-RNA interactions should show a strong signature of negative selection.

To test this expectation, we leveraged the largest pool of human variants from control cohorts released by the Genome Aggregation Database (gnomAD)¹⁸. For each transcribed, noncoding variant in gnomAD (> 20 million SNPs), we interrogated the levels of RBP disruption, specifically using deep learning inference on 232 Seqweaver RBP models (RBP model list Supplementary Table 1). We found significant depletion of strong effect RBP-RNA disrupting variants at high allele frequencies (MAFs, $p < 2.2 \times 10^{-16}$ Wilcoxon rank sum test common vs ultra-rare). The mean variant RBP dysregulation effect increased significantly from common (MAF > 0.05) to ultra-rare inherited variants (MAF < 0.001, $p < 2.2 \times 10^{-16}$ Wald test, Supplementary Fig. 1), which is consistent with a major fitness impact and negative selection acting on variants that disrupt RBP target sites.

Within genes, RBPs frequently exert their regulation via noncoding regions, such as the 5'UTR, introns or 3' UTR. We reasoned that the function of the RBP should be reflected in specific negative selection signatures in the relevant regions. For instance, RBPs that impact mRNA stability would show elevated negative selection signatures in 3'UTRs, whereas splicing RBPs would show these signatures in introns. Importantly, a differential selection signature would shed light on the function of RBPs, and thus providing complementary information to the biochemical studies of RBPs.

To define sub-genic selection signatures, we looked for statistically significant interactions between the genomic location of a variant and the degree of selection acting on RBP dysregulation. We found that diverse RBPs can be segregated by their selection signature across the 5'UTR, 3'UTR or introns (Fig. 1A, 212 RBP models with a Benjamini-Hochberg corrected

FDR < 0.05 sub-genic annotation interaction, full results in Supplementary table 2). The significant association of each RBP with a sub-genic location rules out the baseline interpretation of confounding stochastic non-functional interactions, and points to an active regulatory function with specific fitness consequences.

For many known splicing factors (e.g. SMNDC1 and PRPF8; $p < 2.2 \times 10^{-16}$ Wald test on coefficient of interaction term), intronic variants displayed a significantly elevated level of selection spanning the spectrum of minor allele frequencies compared to the genic background (Fig. 1B). The cap-binding protein NCBP2 showed the most significant enrichment for 5' UTRs, whereas for 3'UTRs, we confirmed the strong human fitness effect of many known mRNA stability, localization and polyA regulatory proteins (e.g. MSI2, PUM2, PABP and ELAVL; all $p < 2.2 \times 10^{-16}$).

Unexpectedly, we found that variants within the 3'UTR that disrupt binding of the superfamily I RNA helicase UPF1 are under a significantly elevated level of negative selection compared to the rest of the gene ($p < 2.2 \times 10^{-16}$). UPF1 is an essential component of the nonsense-mediated decay (NMD) machinery. Our finding is thus unanticipated, but consistent with recent biochemical reports that suggested that UPF1 binding to the 3'UTR can regulate target mRNA stability¹⁹. The regional selection enrichment of UPF1 provides strong corroborating genetic evidence for its role in 3'UTR-mediated post-transcriptional regulation, beyond its canonical NMD function.

Figure 2. Genome-wide RBP dysregulation is a significant source of psychiatric disorder heritability.

a) The per-SNP heritability effect sizes (τ^*) for each RBP target site dysregulation is plotted across the five major psychiatric disorders. The dashed line indicates RBP models below FDR 0.05 threshold after multiple hypothesis correction (block jackknife-based one-sided p-values; Benjamini-Hochberg correction). b) The statistical association between per-SNP heritability and the top psychiatric disorder-associated (TPA) RBPs (mean z scores across the five disorders). The jointly fit representative baseline annotations are also shown for comparison (sub-genic region annotations are highlighted blue e.g. coding region). c) The per-SNP heritability effect sizes (τ^*) for TPA RBPs after conditioning on a collection of molecular QTL annotations (i.e. independent RBP effects from molecular QTLs and baseline annotations). Genotype-Tissue Expression (GTEx), CommonMind, BLUEPRINT (BP). The jointly fit collection of QTL annotation effect sizes is also plotted. Expression (eQTL), splicing (sQTL), mRNA isoform (isoQTL), DNA methylation (mDNA QTL). All error bars are 95% CI.

Variants that disrupt RBP binding is a major contributor to psychiatric disorder risk

Having established the strong importance of RBPs in selection and human fitness, we next investigated the potential role of RBP dysregulation on psychiatric disorder heritability. The strong family history of psychiatric disorders and high heritability estimates make this an

especially important question^{20,21}. To address this hypothesis, we applied the statistical framework of stratified LD score regression²² to partition disease heritability into various functional annotations while directly modeling the extensive LD structure between SNPs. The LD score regression framework allows estimation of SNP effects (τ^*) standardized for comparison across different disease or trait GWAS studies while conditioning on a collection of baseline functional annotations (e.g. coding region, allele age, CpG content, enhancers, promoter and epigenetic histone marks Methods)²³. Here we combine LD score regression with our deep learning framework Seqweaver to estimate the contribution of RBP dysregulation to psychiatric disease.

The stratified LD score regression framework has been tested and shown to produce robust results in large collections of studies²²⁻²⁴. Nevertheless, we performed a comprehensive negative control test in the context of RBP dysregulation. We simulated genetic architecture traits where the underlying causal SNPs were sampled entirely from experimentally profiled brain enhancers, promoters, and brain-expressed protein coding regions (i.e. mostly non-RBP regulatory regions) using real genotypes from the 1000 Genomes Project²⁵ (Methods). Overall, across the 232 RBP models, the simulations produced well-calibrated estimates of RBP dysregulation effect sizes without any upward bias (Supplementary Fig. 2), demonstrating the robustness of our regression models.

Having established the statistical framework, we focused on GWAS from five well-established polygenic psychiatric disorders: ADHD²⁶, autism spectrum disorder²⁷, bipolar disorder²⁸, major depression²⁹ and schizophrenia³⁰. These GWAS were conducted with standardized analysis

pipelines by the Psychiatric Genomics Consortium and therefore minimize the potential sources of technical artifacts. We observed significantly elevated levels of RBP dysregulation effect size (τ^*) estimates across all five psychiatric disorders, with 304 cases where dysregulation of targets of a specific RBP had a significant effect on a psychiatric disorder after correcting for multiple hypothesis testing (Fig. 2A, FDR < 0.05, Benjamini-Hochberg correction, Supplementary Table 3). These results demonstrate a strong underlying causal effect of RBP dysregulation in all five disorders (Fig. 2A).

In addition, we found novel associations between disrupted RBP target sites and RBPs that were previously associated with disease. For example, target sites of the spliceosome-associated factor EFTUD2 contributed significantly across psychiatric disorders (e.g. ADHD $p=1.4 \times 10^{-4}$, SCZ $p=6.4 \times 10^{-4}$ jackknife), and haploinsufficiency of EFTUD2 causes craniofacial malformation, microcephaly and developmental delay⁹, which are shared with many non-Mendelian neurological diseases. In addition, major depression was significantly associated with variants that disrupt target binding by RBFOX, a key splicing regulator in the brain⁶ ($\tau^*=1.4$, $p=8.6 \times 10^{-3}$). Indeed, two GWAS loci within the RBFOX1 RBP gene locus itself are associated with major depression risk²⁹. Overall these data suggest that psychiatric disease is significantly linked with perturbations not only of RBPs, but also their targets, which represent a much larger set of variants spread across the transcribed regions of the genome.

RBP target effects exceed those of coding variants and broadly impact post-transcriptional regulation

Next, we examined how variants that dysregulate RBP function compare to other functional variant categories, by comparing across the jointly fit annotations in the regression models. We

found that the statistical association between disease heritability and RBPs are among the top functional annotations, exceeding the collective set of coding variants or previously annotated epigenetic regions (Fig. 2B). Thus, noncoding variants that perturb RBP binding significantly increase psychiatric disorder risk. Furthermore, within gene regions, the collective liability of RBP dysregulation can exceed coding region variant effects (Supplementary Fig 3). These results remain significant and robust after conditioning on potential confounding factors, such as background selection rate, low levels of LD (LLD), allele age and minor allele frequency (Fig. 2B baseline annotations included in the regression model, Methods).

Notably, the strong noncoding dysregulation effects that we identified include diverse categories of RBPs, beyond splicing factors. For example, the most well-powered of the five GWAS, schizophrenia, 49/91 significant RBP models ($FDR < 0.05$) were UTR regulatory RBPs. This observation reveals a wider importance of post-transcriptional effects beyond splicing, and was not limited to schizophrenia: the top psychiatric disorder-associated (TPA) RBPs (mean z-score > 2.5) covered all the major RNA regulatory modes from splicing to transcript stability, based on the shared risk across all five psychiatric disorders studied (Fig. 2B). For example, UPF1¹⁹ and FAM120a³¹, which regulate transcript degradation, showed consistent strong signals across the disorders, with top ranked effect sizes in schizophrenia (UPF1 $\tau^* = 1.16$ $p = 2.0 \times 10^{-6}$, FAM120a $\tau^* = 1.2$ $p = 1.4 \times 10^{-7}$). We also observed a pair of ATP-dependent RNA helicases (DDX6, DDX42) among the TPA RBPs, adding support to the neuropathogenic role of helicases³². Overall, these data demonstrate that disruption of diverse types of post-transcriptional regulation are highly associated with psychiatric disease burden.

Finally, to replicate our findings of RBP dysregulation effects in an independent cohort, we utilized cross-psychiatric disorder GWAS summary statistics released by the iPSYCH consortium³³. Risk variants identified in this study largely represent loci with shared underlying genetic risk factors across psychiatric disorders, and are based on a homogenous Danish population diagnosed using the same public healthcare system criteria. We focused this replication analysis on schizophrenia, which had no overlapping individuals within the Psychiatric Genomics Consortium and iPSYCH cohorts. We found highly significant concordance of RBP-associated risk between the two cohorts ($P < 2.2 \times 10^{-16}$ Wilcoxon rank sum test, see Supplementary Fig. 4 for RBP dysregulation effect sizes). Thus, our RBP dysregulation estimates are robust and not a by-product of population stratification.

RBP target effects explain substantial heritability beyond known molecular QTLs

Previous reports have found that molecular QTLs are strongly enriched for disease heritability^{13,34}. Thus we investigated whether the profiled RBP dysregulation effects capture information about disease that is independent of the large-scale molecular QTL studies. We estimated the effect size of each RBP for each disorder while jointly conditioning on the molecular QTL-based annotations from GTEx²⁴, CommonMind³⁵ and BLUEPRINT³⁶ consortium (in addition to all baseline annotations). We found that TPA RBPs remain highly significant and display overall greater effect sizes compared to the QTL annotations (Fig. 2C, Supplementary Table 4). Importantly, this implies that RBP dysregulation effects are largely independent from known molecular QTLs and should provide an important additional tool for dissecting genetic architectures of disease.

Figure 3. RBP dysregulation underlies shared and distinct genetic architectures of psychiatric phenotypes

a) The meta-analysis of RBP dysregulation effect sizes in loss-of-function (LoF) intolerant genes compared to the remaining set of genes is shown (95% CI). All effect sizes are conditioned on baseline annotations plus two additional annotations delineating LoF intolerant gene and their coding regions to control for the higher background heritability rates. b) Top psychiatric disorder-associated (TPA) RBP estimated effect sizes across a collection of psychiatric traits and non-brain related phenotypes are shown (95% CI). The adjacent heatmap displays spearman rank

correlation for all RBP effect sizes between the phenotype and psychiatric disorders (mean effect). c) RBP dysregulation effect sizes (τ^*) for differential risk between bipolar disorder and schizophrenia, estimated for both all genes and for LoF intolerant genes. The dashed line indicates RBP models below FDR 0.05 threshold (Benjamini-Hochberg method).

RBP dysregulation contributes to shared and distinct aspects of psychiatric disorders

Having established the significant role of RBP dysregulation in disease, we next examined if altered RBP-RNA interactions underlie the extensively shared genetic architectures across psychiatric phenotypes. At the gene level, psychiatric disorder heritability is enriched for mutation-intolerant genes³⁷. Here we investigated if this enrichment could be in part explained by variants that impact RBP regulation. Indeed, we observed that RBP effect sizes were significantly larger for target site variants within loss-of-function (LoF)-intolerant genes (Fig. 3A, $P < 2.2 \times 10^{-16}$ paired Wilcoxon rank sum test for RBP effect size τ^* , LoF intolerant defined by ExAC³⁸ and controlled for different baseline gene heritability enrichment levels Methods).

Moreover, we found that numerous RBP models were significantly associated with heritability for a broader set of psychiatric and related phenotypes examined by the Brainstorm Consortium²⁴ (Supplementary Fig. 5, 856 RBP model-phenotype pairs FDR < 0.05 with Benjamini-Hochberg correction). This association was especially strong for psychiatric traits like “Cigarettes per day” (a common proxy for addictive risk-taking behavior³⁹) and “Depressive symptoms” (a widely shared clinical feature for many psychiatric disorders⁴⁰), whereas the non-brain related phenotypes displayed lower correlation of overall RBP effect sizes with psychiatric disorders (Fig. 3B). These results imply that variants disrupting RNA-RBP interactions affect neuropathogenic pathways and are a significant driver of the high genetic correlations observed between psychiatric disorders and traits.

On the other hand, we found that the disruption of distinct RBP target sites can help explain differences between psychiatric disorders. Specifically, we found that disruption of targets for the RBP ILF3 contributes to the differential liability between schizophrenia and bipolar disorder (Fig. 3C, two biological replicate ILF3 models highlighted, $p=4.1 \times 10^{-5}$ LoF intolerant genes, jackknife), extending findings of the Psychiatric Genomics Consortium study²⁸. Moreover, at the gene-level, ILF3 was the 5th most significantly associated gene in bipolar disorder (Supplementary Fig. 6A, $p=1.2 \times 10^{-9}$, MAGMA statistical framework⁴¹, colocalization analysis further supports this ILF3 association Supplementary Fig. 6B), but it had no significant association in the better-powered schizophrenia study³⁰. Our analysis suggests that the molecular network composed of both RBP ILF3 and its targets helps to discern these two psychiatric disorders. Determining how the ILF3 network alters brain cellular functions will shed light on how genetics influence clinical outcome variation.

Figure 4. Functional RBP regulatory mapping identifies a schizophrenia risk variant in DDHD2 3'UTR.

a) Schizophrenia GWAS signal for the cross-ethnic associated region. The highlighted SNP rs6981405 represents the top predicted RBP dysregulation variant disrupting RBP QKI binding. b) ENCODE eCLIP data confirms SNP rs6981405 C>A leads to the disruption of RBP QKI binding to DDHD2 (rs6981405 genotype for cell line K562 homozygous AA, HepG2 homozygous CC). SNPs were allowed during CLIP read alignment. c) QKI knockdown coupled by RNA-seq confirms the QKI-mediated regulation of DDHD2, which is disrupted in the homozygous AA genotype (i.e. QKI KD shows no effect when SNP rs6981405 impedes with RBP binding). Error bars represent SEM. d) Schematic: the variant at rs6981405 disrupts the QKI - DDHD2 3'UTR interaction, which alters the abundance of mature DDHD2 mRNA, and, in turn, schizophrenia risk.

Functional mapping based on RBP dysregulation identifies DDHD2 as a schizophrenia risk gene

Hundreds of genomic regions are associated with the risk of psychiatric disorders, consistent with a polygenic architecture³⁰. However, very few disease-associated regions have been mapped to their causal SNPs, with the underlying biochemical mechanism dissected. As a case study, we leveraged our ability to interrogate allelic-specific RBP target site dysregulation genome-wide to investigate a schizophrenia risk region.

The 8p12 genomic region was first identified as a significant schizophrenia risk region in the Han Chinese population⁴² and was subsequently found significantly linked in Europeans³⁷. Cross-population replication implies a robust molecular cause underlying the associated loci with global clinical prospective. Our analysis provides a potential biochemical mechanism for this association: within this region we identified a SNP in the DDHD2 3'UTR that can disrupt binding by the RBP QKI, which is known to play an important role in schizophrenia^{43,44} (Fig. 4A, this top Seqweaver predicted SNP rs6981405 was a fine-mapped candidate SNP (95% credible set)³⁷).

ENCODE⁴⁵ QKI eCLIP data, available in two cell lines (K562 and HepG2), support this association. The candidate SNP rs6981405 is homozygous CC in HepG2 and homozygous AA in K562. QKI and its target DDHD2 are highly expressed in both cell lines (> 15 TPM). Importantly, QKI-DDHD2 binding is observed only in homozygous C allele genotype (Fig. 4B), consistent with our estimation (i.e. A allele disruption of QKI binding). Furthermore, RNAi-mediated depletion of QKI led to elevated levels of DDHD2 mRNA in C allele genotype, but not

in homozygous A allele genotype line, where QKI binding is already disrupted (Fig. 4C). Thus, mutation of this SNP in DDHD2 mRNA precludes its regulation by QKI.

DDHD2 is a principal brain triglyceride lipase, that when mutated causes a hereditary neurological disease, spastic paraplegia⁴⁶. Our genetic evidence, provided by RBP regulatory mapping, coupled with supporting experimental data, suggest that QKI-mediated regulation of DDHD2 transcript levels influences the risk of schizophrenia and imply a pathogenic role for altered lipid metabolism in this disease.

Discussion

A critical challenge in human disease research involves moving from cataloging disease risk loci to understanding the molecular mechanism. Interactions between transcripts and RBPs are early events that influence protein expression and function. Therefore, interrogating genetic architectures at this early layer of molecular regulation presents a powerful opportunity, as it reduces the complexity of having to deconvolute disease correlated signal into causal factors that often burdens the analysis for further downstream approaches. Most importantly, targeted biochemical perturbation of RBP-RNA interactions has proven to be an effective new avenue of clinical intervention⁴⁷, and therefore delineating RBP target sites as a major source of molecular dysfunction contributing to psychiatric disorders is a vital task.

In this work, we establish that RBP dysregulation is a key contributor to human fitness by identifying extensive negative selection signatures in the largest-to-date WGS gnomAD¹⁸ cohort. We further find that concentrated regional fitness effects observed for each RBP provide a

genetic indicator for the underlying biochemical regulatory function. We also highlight that disruption of diverse RBP functions significantly affect fitness, supporting an extensive pathogenic contribution beyond splicing regulation.

Next, focusing on psychiatric disorders, we provide support for a significant causal role for RBP dysregulation, linking inherited risk variants to biochemical perturbations that ultimately lead to psychiatric clinical phenotypes. Intriguingly, one key theme we find that emerged is the psychiatric risk convergence at both the RBP regulatory factor (i.e. RBP protein itself) and its downstream target site dysregulation. For instance, variants within RBFOX and its downstream targets are linked to major depression risk, and variants within EFTUD2 and its downstream targets are linked to neurological dysfunction. In addition, we find converging evidence links RBP ILF3 and its RNA targets to the molecular differences between schizophrenia and bipolar disorder. Similarly, the RBP “fragile X mental retardation protein” (FMRP) is the most common monogenic cause of autism⁴⁸, and FMRP mRNA targets are highly linked to both autism and schizophrenia⁴⁹⁻⁵¹. Thus, these converging RBP regulatory networks may present ideal clinical targets, due to the greater collective biochemical contribution to pathogenicity.

Methodologically, we demonstrate that our deep learning inference of genome-wide molecular effects allows us to estimate major modes of biochemical perturbation and their contribution to disease. We find that splicing disruption is the tip of the iceberg, as we discovered widespread psychiatric disease risk associated to RBPs that regulate each step of the RNA life cycle from processing to localization to degradation. Current molecular QTL resources, while incredibly valuable, lack the breadth to capture these diverse molecular functions of RBPs. This caveat

limits the scope of analysis for disease, now encapsulating hundreds of thousands of cohort samples (e.g UK biobank⁵²). Indeed, there has been increasing evidence of an extensive pathogenic role of RBPs from cancer⁵³, autoimmune disease^{54,55} to myopathy⁵⁶. Our computational framework can enable studies of RBP dysregulation in these and other disorders on the whole-genome-scale. Furthermore, as more tissue- and cell-type-specific CLIP data becomes available, this approach can provide a data-driven window into tissue-specific RBP dysregulation in disease.

To enable rapid analysis of psychiatric diseases and the extension to the greater collection of disease GWAS studies, we now profiled and made available genome-wide inference of RBP target site dysregulation effects for the largest collection of human variation identified by the gnomAD cohort. This resource, with the entire spectrum of common to ultra-rare variants, should provide the means to interrogate RBP-derived human diseases at an unprecedented scale.

Methods

Deep learning inference of RBP dysregulation variant effects

We utilized the machine learning approach of deep convolutional neural networks (CNN)⁵⁷ to build a quantitative model of the RNA sequence features required for each RBP binding biochemically assayed by CLIP⁵⁸ (training data). These RBP models subsequently enable the probabilistic inference of variant impact, capturing both direct and indirect effects, on the RBP

binding potential. The applied Seqweaver RBP model architecture and training are described in our previous *de novo* mutation autism work¹⁶, and to ensure that our current results can be directly comparable, we used the identically 232 CLIP-based RBP models (Supplementary Table 1) without any modifications (i.e. no additional training or manual parameter changes).

Briefly, CNNs allow researchers to design network architectures that can leverage information of high order motifs at different spatial scales but with optimal parameter sharing to avoid overfitting. Our Seqweaver DNN architecture consists of an initial input layer followed by a series of convolution and pooling layers. The input sequence layer contains a 4 x 1,000 matrix that encodes the input RNA sequence of U, A, G, C across the 1,000 bp window anchored around the RBP binding site. The subsequent convolution layer looks at 8 bps at a time shifting by 1 bp and computes the convolution operation of 160 kernels. At this first convolution level, the kernels are equivalent to searching for a collection of local sequence motifs in a one-dimensional RNA sequence. Analogues to neurons, we then apply a rectifier activation function (ReLU) that sets the convolution layer output to a scale of minimum of 0 (i.e. $\text{ReLU}(x) = \max(0, x)$). Thus formally, input S results in convolution layer output location n for kernel k as the following:

$$\text{Convolution}(S)_{n,k} = \text{ReLU} \left(\sum_i^I \sum_d^D w_{i,d}^k S_{n+i,d} \right)$$

where I is the window size and D is the input depth (e.g. for the first convolution layer D represents the four RNA bases).

Next, we add a pooling layer that allows the reduction of the dimensional size of the network and parameters. Specifically, every window of 4 for a kernel output are collapsed into the maximum

value observed in that span. Subsequently, the resulting output is used as input for a sequence of convolution (2nd), ReLU, pooling and convolution layer (3rd) in which higher order sequence motifs can be derived based on the first layer local motifs (2nd conv. layer 320 kernels, 3rd conv. layer 480 kernels with identical ReLU and pooling layer).

Lastly, we add a fully connected layer that can now take the resulting output from the three convolution steps to integrate across the entire 1,000 bp context to derive a final set of higher order sequence motifs. These sequence motifs are shared across all RBP models that allow optimal parameter reduction, but also are based on the biological intuition that many RNA sequence features are shared in the cell (e.g. RNA polyA signal, splice sites and branchpoints). The fully connected layer outputs are then subjected to RBP-specific weighted logistic functions (sigmoid, [0,1] scale) allowing for the simultaneous prediction of each RBP binding potential to the input RNA sequence.

Finally, for variant effect prediction, we take the absolute predicted probability differences between the two alleles (reference vs. alternative) computed by the convolutional neural network for each 232 RBP models. Important to note that no variant-level sequence information was used during the training of our Seqweaver RBP models, therefore we are not limited or biased by any variant-level training set. The final RBP variant effects were set to [0,1] scale for fitting LD score regression models. All training data and RBP models can be downloaded at hb.flatironinstitute.org/Seqweaver.

Genome-wide RBP dysregulation analysis of negative selection

The 2.1 release of gnomAD¹⁸ cohort variants, all passing the random forest gnomAD quality filter were downloaded and filtered for noncoding region variants (i.e. nonrepeat regions of 5'UTR, intron and 3'UTR) of protein coding genes (we use AC >1 to filter for inherited variants). The resulting final 21,513,861 SNP variants were used in the analysis.

For each Seqweaver RBP model, the distribution of absolute predicted probability differences (ref. vs alt. allele) across all variants were standardized to have a standard deviation 1 to obtain the final RBP dysregulation estimates. The gnomAD cohort allele frequencies were used to segregate the variants into four different minor allele frequency bins (>0.05, 0.05~0.01, 0.01~0.001, <0.001) and then to obtain the mean variant RBP dysregulation levels.

Variant level annotations to sub-genic regions were conducted as previously described¹⁶, annotating to 5'UTR, 3'UTR or 200bps introns flanking an exon previously observed to be alternatively spliced⁵⁹. The RBP sub-genic selection signature was assessed by fitting a linear model regressing RBP dysregulation levels on allele frequency and sub-genic annotations, and evaluating the statistical significance of the interaction term between the two explanatory variables. That is querying for statistically significant interactions between a variants gene location and the degree of selection acting on RBP target site dysregulation. For downstream analysis, UTR regulatory RBPs were defined by RBP models that showed sub-genic selection signatures with FDR < 0.05 for only 3' or 5' UTR regions.

Estimating the RBP dysregulation GWAS effect sizes

The extensive linkage disequilibrium (LD) between SNPs in the human population provide an analytical challenge for estimating the true underlying effect size for RBP dysregulation from GWAS. For example, high χ^2 statistic SNPs in the 3'UTR may appear to be an indication of UTR mediated risk to a disease, but conversely may simply just be tagging the true causal enrichment of protein coding region SNPs due to the high LD in the region. To resolve this challenge, we use the previously published statistical framework of stratified LD score regression²² to estimate the RBP dysregulation effect sizes for each examined trait or disease GWAS. More specifically from the summary statistics of a GWAS, we can write the expected χ^2 value for SNP j as

$$E[\chi_j^2] = N \sum_c \tau_c l(j, c) + Nb + 1$$

where N is sample size and the annotation specific “LD score” $l(j, c)$, representing annotation (c)’s cumulative effects tagged by the SNP j , can be written as

$$l(j, c) = \sum_k a_c(k) r_{jk}^2$$

where $a_c(k)$ is the annotation value at SNP k (e.g RBP dysregulation level or coding SNP), r_{jk} is the correlation between SNP j and k in the reference panel (selected to best match the GWAS cohort), and b measuring the confounding bias²². Lastly, τ_c and the final standardized form τ_c^* – normalized by the total SNP-based heritability and s.d. of an annotation – represents the estimated effect size of the annotation²³.

$$\tau_{rbp}^* = \frac{M * sd_{rbp}}{h^2} \tau_{rbp}$$

More formally for RBP dysregulation annotations, τ^* represents the per-SNP heritability ($\frac{h^2}{M}$, M number of common SNPs) associated with a standard deviation increase of variant RBP effect

(sd_{rbp}). We restrict our RBP predictions to SNPs from 1000 Genomes Project (European cohort), and fit the regression model only on HapMap SNPs with MAF > 0.05 as previously conducted²². Block jackknife procedure is used to test the statistically deviation from zero for each fitted τ^* .

As presented in the regression model, we fit τ_{rbp}^* by conditioning on a collection of baseline annotations to avoid upward bias in the effect size estimation. We obtain the collection of baseline annotations previously used in the stratified LD score regression study (i.e. baselineLD)²³, that includes functional annotations such as coding regions, 3'/5' UTR, intron, promoter, transcription start site (TSS) and multiple epigenetic marks. We included a new functional baseline annotation that labels all gene region SNPs, controlling for baseline effects tagging transcribed regions, that collectively results in appropriately calibrated null uniform RBP p-values based on permutation test shown in Supplementary Fig. 7. Additional baseline annotations include non-functional annotations such as allele age, minor allele frequency, low levels of LD, CpG content and background selection statistics. We excluded conservation-based annotations, as RBP regulatory binding sites are known to be highly conserved⁶⁰⁻⁶². The final reported RBP effect size (τ_{rbp}^*) were jointly fit, iteratively for each RBP, with all baseline annotations (full 71 baseline annotations listed in Supplementary Table 5).

Simulations for RBP effect size estimation

We conducted simulations to ensure that our regression models produced unbiased RBP effect sizes. Specifically, we tested to verify that false positive results were not obtained for genetic architectures where the causal SNPs were derived from functional elements that are largely non-

RBP regulatory regions – epigenetic enhancers, promoters and protein coding regions. We simulated 400 GWAS summary statistics using the 1,000 Genomes Project European reference panel using simGWAS⁶³. Testing for two scenarios, in each simulation, we sampled 1% or 5% SNPs (MAF > 0.01 and chromosome 1), as the causal set from brain epigenetic enhancers annotated by the PsychENCODE Consortium⁶⁴, and both promoters⁶⁵ and coding regions (restricted to nonsynonymous variants) that are expressed in the brain⁶⁶. For each causal SNP effect size, we model as a Fisher polygenic model with trait heritability set to $h^2 = 0.5$. Each simulated GWAS was fit with our LD score regression model (RBP + baseline annotations) to obtain the RBP effect size estimate (τ_{rbp}^*). Results for the simulations produced overall robust unbiased estimates across our RBP models (Supplementary Fig. 2).

GWAS disease and trait selection

We selected psychiatric disorder GWAS studies conducted by the Psychiatric Genomics Consortium, that was uniformly processed and analyzed, and was sufficiently powered to observe genome-wide significant SNPs, resulting in five disorders – ADHD, autism spectrum disorder, bipolar disorder, major depression and schizophrenia. To facilitate cross-study comparison, we selected psychiatric traits and non-brain associated diseases previously profiled by the Brainstorm consortium study (excluded traits that did not find genome-wide significant SNPs). The Danish cohort from iPSYCH consortium cross-disorder GWAS study³³ was used as replication analysis. Population matched schizophrenia case vs. bipolar disorder case GWAS²⁸ summary statistic was obtained from the Psychiatric Genomics Consortium web portal (full list of GWAS studies examined in this work Supplementary Table 6).

Joint modeling of molecular QTLs

Fine-mapped GTEx²⁴ eQTL (FE-meta) and BLUEPRINT⁴⁵ molecular QTL annotations were obtained from a previous study³⁴, that generated and validated the max causal posterior probability (MaxCPP)-based QTL annotations for GWAS enrichment analysis. The CommonMind⁶⁷ isoformQTLs and GTEx sQTL (brain cortex version 8) were fine-mapped to produce MaxCPP annotations following the same procedure as was previously reported³⁴. The collection of molecular QTL MaxCPP annotations and all baseline annotations were jointly modeled in the stratified LD score regression with each RBP annotation to estimate the disease associated effect sizes.

Genetic architecture analysis

Loss-of-function intolerant genes were obtained by the ExAC consortium with pLI threshold of > 0.9 as previously described¹⁶. For stratified LD score regression models, we jointly fit, for each RBP model, the LoF intolerant gene and non-LoF intolerant gene regions variant RBP effect sizes (τ^*) by splitting the RBP annotation into two by gene set. We added two additional baseline annotations for this analysis, that includes SNP to LoF intolerant gene regions, and SNP to LoF intolerant gene's coding region, to prevent potential upward bias due to the general higher background heritability enrichment levels. We also added the two SNP to LoF intolerant gene or their coding region baseline annotations for the differential risk analysis between schizophrenia and bipolar disorder to mitigate any potential bias.

MAGMA⁴¹ was used to estimate the gene level association with schizophrenia and bipolar disorder. GENCODE⁶⁸ v25 gene annotations lifted to GRCh37 coordinates and total 19,984

protein coding genes were analyzed. We used SNPs from upstream 10k, gene body and downstream 1.5k for each gene as was previously used in a Psychiatric Genomics Consortium schizophrenia GWAS analysis study⁶⁹ along with the 1000 Genomes Project²⁵ European reference panel. Colocalization analysis for the ILF3 locus between schizophrenia vs. bipolar GWAS²⁸ and GTEx ILF3 eQTL (v8 meta-tissue) was conducted using Coloc⁷⁰.

Functional mapping of DDHD2

QKI eCLIP and knockdown RNA-seq data was obtained by the ENCODE project⁴⁵ in K562 and HepG2 cell lines. eCLIP data was processed as previously described¹⁶ and visualized in IGV⁷¹. Kallisto⁷² coupled with Sleuth⁷³ was used for differential expression analysis of DDHD2 transcript (ENST00000520272) following QKI KD. P-values were computed using likelihood ratio test implemented in Sleuth and FDR was computed across all transcripts using Storey's q-value method⁷⁴. Genotyping results for SNP rs6981405 in K562 and HepG2 lines were obtained from ENCODE project.

Code and data availability

The code is available from <https://hb.flatironinstitute.org/seqweaver/about>. All variant predicted scores have been made available to download and as an interactive web interface available at <https://hb.flatironinstitute.org/seqweaver>.

References

1. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50**, 1112-1121 (2018).
2. Black, D.L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry* **72**, 291-336 (2003).
3. Krichevsky, A.M. & Kosik, K.S. Neuronal RNA Granules: A Link between RNA Localization and Stimulation-Dependent Translation. *Neuron* **32**, 683-696 (2001).
4. Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11-23 (2001).
5. Costa-Mattioli, M., Sossin, W.S., Klann, E. & Sonenberg, N. Translational Control of Long-Lasting Synaptic Plasticity and Memory. *Neuron* **61**, 10-26 (2009).
6. Darnell, R.B. RNA Protein Interaction in Neurons. *Annual Review of Neuroscience* **36**, 243-270 (2013).
7. Schuman, E.M. mRNA Trafficking and Local Protein Synthesis at the Synapse. *Neuron* **23**, 645-648 (1999).
8. Verkerk, A.J.M.H. *et al.* Identification of a gene FMR-1 containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905-914 (1991).
9. Lines, Matthew A. *et al.* Haploinsufficiency of a Spliceosomal GTPase Encoded by EFTUD2 Causes Mandibulofacial Dysostosis with Microcephaly. *The American Journal of Human Genetics* **90**, 369-377 (2012).
10. Bernier, Francois P. *et al.* Haploinsufficiency of SF3B4, a Component of the Pre-mRNA Spliceosomal Complex, Causes Nager Syndrome. *The American Journal of Human Genetics* **90**, 925-933 (2012).
11. Messiaen, L.M. *et al.* Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Human Mutation* **15**, 541-555 (2000).
12. Xiong, H.Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
13. Li, Y.I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600 (2016).
14. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
15. Walker, R.L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179**, 750-771.e22 (2019).
16. Zhou, J. *et al.* Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature Genetics* **51**, 973-980 (2019).
17. Power, R.A. *et al.* Fecundity of Patients With Schizophrenia, Autism, Bipolar Disorder, Depression, Anorexia Nervosa, or Substance Abuse vs Their Unaffected Siblings. *JAMA Psychiatry* **70**, 22-30 (2013).
18. Karczewski, K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).

19. Hogg, J.R. & Goff, S.P. Upfl Senses 3'UTR Length to Potentiate mRNA Decay. *Cell* **143**, 379-389 (2010).
20. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
21. Laursen, T.M. *et al.* Family History of Psychiatric Illness as a Risk Factor for Schizoaffective Disorder: A Danish Register-Based Cohort Study. *Archives of General Psychiatry* **62**, 841-848 (2005).
22. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-1235 (2015).
23. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421-1427 (2017).
24. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
25. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
26. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics* **51**, 63-75 (2019).
27. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics* **51**, 431-444 (2019).
28. Ruderfer, D.M. *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715.e16 (2018).
29. Wray, N.R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics* **50**, 668-681 (2018).
30. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
31. Kelly, T.J., Suzuki, H.I., Zamudio, J.R., Suzuki, M. & Sharp, P.A. Sequestration of microRNA-mediated target repression by the Ago2-associated RNA-binding protein FAM120A. *RNA* **25**, 1291-1297 (2019).
32. Balak, C. *et al.* Rare De Novo Missense Variants in RNA Helicase DDX6 Cause Intellectual Disability and Dysmorphic Features and Lead to P-Body Defects and RNA Dysregulation. *The American Journal of Human Genetics* **105**, 509-525 (2019).
33. Schork, A.J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nature Neuroscience* **22**, 353-361 (2019).
34. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature Genetics* **50**, 1041-1047 (2018).
35. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19**, 1442-1453 (2016).
36. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
37. Pardiñas, A.F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* **50**, 381-389 (2018).
38. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

39. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics* **51**, 245-257 (2019).
40. Howard, D.M. *et al.* Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nature Communications* **9**, 1470 (2018).
41. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
42. Shi, Y. *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nature genetics* **43**, 1224-1227 (2011).
43. Åberg, K. *et al.* Human QKI, a new candidate gene for schizophrenia involved in myelination. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **141B**, 84-90 (2006).
44. Bhalala, O.G., Nath, A.P., Consortium, U.K.B.E., Inouye, M. & Sibley, C.R. Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. *PLOS Genetics* **14**, e1007607 (2018).
45. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
46. Inloes, J.M. *et al.* The hereditary spastic paraplegia-related enzyme DDHD2 is a principal brain triglyceride lipase. *Proceedings of the National Academy of Sciences* **111**, 14924 (2014).
47. Finkel, R.S. *et al.* Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *New England Journal of Medicine* **377**, 1723-1732 (2017).
48. de la Torre-Ubieta, L., Won, H., Stein, J.L. & Geschwind, D.H. Advancing the understanding of autism disease mechanisms through genetics. *Nature Medicine* **22**, 345-361 (2016).
49. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
50. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
51. Darnell, Jennifer C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261 (2011).
52. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).
53. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell* **156**, 1324-1335 (2014).
54. Darnell, R.B. & Posner, J.B. Paraneoplastic Syndromes Involving the Nervous System. *New England Journal of Medicine* **349**, 1543-1554 (2003).
55. Häslér, R. *et al.* Alterations of pre-mRNA splicing in human inflammatory bowel disease. *European Journal of Cell Biology* **90**, 603-611 (2011).
56. Cummings, B.B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**, eaal5209 (2017).
57. Krizhevsky, A., Sutskever, I. & Hinton, G.E. *ImageNet classification with deep convolutional neural networks*, 84-90 (Association for Computing Machinery, 2017).

58. Licatalosi, D.D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469 (2008).
59. Yan, Q. *et al.* Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proceedings of the National Academy of Sciences* **112**, 3445 (2015).
60. Sorek, R. & Ast, G. Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse. *Genome Research* **13**, 1631-1637 (2003).
61. Lebedeva, S. *et al.* Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Molecular Cell* **43**, 340-352 (2011).
62. Lovci, M.T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural & Molecular Biology* **20**, 1434-1442 (2013).
63. Fortune, M.D. & Wallace, C. simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics. *Bioinformatics* **35**, 1901-1906 (2018).
64. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
65. Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554-566 (2015).
66. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).
67. Hoffman, G.E. *et al.* CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Scientific Data* **6**, 180 (2019).
68. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774 (2012).
69. Skene, N.G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics* **50**, 825-833 (2018).
70. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).
71. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178-192 (2012).
72. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525-527 (2016).
73. Pimentel, H., Bray, N.L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods* **14**, 687-690 (2017).
74. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440 (2003).