HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2021 July 18.

Published in final edited form as:

Nat Genet. 2021 February ; 53(2): 166–173. doi:10.1038/s41588-020-00761-3.

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

Christopher Y. Park[1,*], **Jian Zhou**[1], **Aaron K. Wong**[1], **Kathleen M. Chen**[1], **Chandra L. Theesfeld**[2], **Robert B. Darnell**[3,4], **Olga G. Troyanskaya**[1,2,5,*]

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

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

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

[4]Howard Hughes Medical Institute

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

Abstract

Despite the strong genetic basis of psychiatric disorders, the underlying molecular mechanisms remain 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. However, quantifying the pathogenic contribution of noncoding variants impacting RBP target sites is challenging. Here, we leverage a deep learning approach that can accurately predict RBP target site dysregulation effects of mutations, and discover that RBP dysregulation is a principal contributor to psychiatric disorder risk. RBP dysregulation explains a substantial amount of heritability not captured by large-scale molecular QTL studies, and has a stronger impact than common coding region variants. We share genome-wide profiles of RBP dysregulation, which we use to identify *DDHD2* as a candidate schizophrenia risk gene. This resource provides a novel analytical framework to connect the full range of RNA regulation to complex disease.

Interrogating the genetics underlying psychiatric disorders is a key path to understanding the pathophysiological cause of mental illness. In particular, genome-wide association studies

(*)Corresponding authors: Olga G. Troyanskaya, ogt@cs.princeton.edu, Christopher Y. Park, cpark@flatironinstitute.org.

Author contributions

C.Y.P. and O.G.T. conceived the study. C.Y.P. designed the study, developed the computational methods and performed the analyses. J.Z. and C.L.T. contributed ideas and insights. A.K.W. and K.M.C. developed the web interface/software. C.Y.P. and O.G.T. wrote the manuscript.

Competing Interests

None declared.

(GWAS) have become a widely adopted approach for studying the genetics of human psychiatric disorders. With ever-increasing sample cohorts, approaching one million subjects for psychiatric disorders[1], numerous risk loci have now been cataloged. Despite this progress, the biochemical perturbations and pathological mechanisms underlying human psychiatric disorders remain challenging to decipher, critical as this is for translating genetic discoveries into sorely needed actionable targets.

RNA-binding proteins (RBP) regulate various aspects of RNA metabolism, including RNA splicing[2], localization[3], stability[4] and translation[5]. Each of these functions are critical not only for protein expression, but also for proper spatiotemporal function, especially in the brain[6]. Proper regulation across the full RNA life cycle by RBPs is critical in neurobiology, where complex regulatory events take place in synapses far away from the nucleus[7].

RBP-encoding genes are frequently mutated in neuropsychiatric disorders, suggesting a pathogenic role[8–10] and inspiring follow-up efforts to study the larger set of variants at the RNA target sites (*trans*) that impact RBP-RNA interactions[11–15]. These studies have shed light on the roles of specific RBPs and functions, particularly splicing, in mental disorders. However, comprehensive, genome-level insights that span diverse RBPs, their biochemical functions and casual target sites are lacking. Thus, the extent to which *trans* dysregulated RNA-RBP interactions contribute to psychiatric disorders remains an open question.

In this study, we address this challenge with the first genome-wide, systematic analysis of the role of RBP target site dysregulation in psychiatric disease (i.e. the impact of dysregulation at *trans*-regulatory RBP targets). To map the impact of variants on RBP-RNA interactions at scale, we leverage a deep learning-based sequence model, Seqweaver, whose accuracy in predicting RBP target site dysregulation we previously extensively evaluated both computationally and experimentally, applying it to detect *de novo* noncoding mutation signal in autism in probands versus their unaffected siblings[16]. Here, we use Seqweaver to build an unprecedented profile of allele-specific effects of inherited variants genome-wide, enabling us to examine the diverse landscape and impact of RBP dysregulation in complex psychiatric disorders.

Our study generates genome-wide annotations of variants linked to RBP dysregulation, publicly available at hb.flatironinstitute.org/seqweaver. We show that across RBPs with diverse functions, the dysregulation of RBP target sites are top drivers of psychiatric disorder risk. Leveraging this resource, we discover a novel link between a RBP-disrupting variant in *DDHD2*, a phospholipase involved in hereditary spastic paraplegia, and a multiethnic-associated locus that increases the risk of schizophrenia. Our study provides an analytical framework that will greatly facilitate and accelerate biochemical investigation of variants linked to complex disorders.

Results

RBP function is reflected in genome-wide negative selection signatures

A substantial gap still exists between connecting variants linked to RBP target site dysregulation and subsequent phenotypic consequences. We previously built a quantitative

model, Seqweaver[16], that can accurately estimate RBP binding based on sequence information (Supplementary Fig. 1), trained on *in vivo* RBP-RNA interaction profiles (CLIP[17] experiments). The single-nucleotide sensitivity in our model allows changes in RBP target site binding for any given variant to be quantified (defined as RBP dysregulation effects), and thus individual RBP phenotypic effects to be mapped to these genetic changes. Because our models enable exploration of RBP dysregulation at a genome-wide scale, we first sought to analyze selection traces that act on binding site sequences that would impact the RBP regulation. Case studies exploring pathogenic *de novo* mutations[18] predict that negative selection is expected to act strongly on large effect mutations. We hypothesized that RBP target site variants that lead to dysregulation are also subject to negative selection.

To test this expectation, we leveraged the largest pool of human variants from control cohorts released by the Genome Aggregation Database (gnomAD)[19]. For each transcribed, noncoding variant in gnomAD (> 20 million SNPs), we interrogated the levels of RBP dysregulation using deep learning inference based on 232 Seqweaver RBP models (RBP model list Supplementary Table 1). We found significant depletion of strong effect RBP dysregulation variants at high allele frequencies (MAF > 0.05, $p < 2.2 \times 10^{-16}$ Wilcoxon rank sum test common vs ultra-rare (MAF <0.001)). The mean impact of target site variants increased significantly from common (MAF > 0.05) to ultra-rare inherited variants (MAF < 0.001, $p < 2.2 \times 10^{-16}$ Wald test, Extended Data Fig. 1), which is consistent with the disruption of RBP target sites having a major impact on fitness, and negative selection acting on such variants.

Within genes, RBPs frequently exert their regulatory role by acting on noncoding regions, such as the 5'UTR, introns or 3' UTR. We reasoned that RBP function should be reflected in specific negative selection signatures in these noncoding regions. For instance, RBPs that impact mRNA stability would show elevated negative selection signatures in 3'UTRs, whereas signatures for splicing RBPs should be found in introns. Importantly, localization of selection signatures to specific gene regions would shed light on RBP function, and provide information complementary to 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 whole 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 $p < 2.2\times10^{-16}$).

Unexpectedly, we found that 3'UTR variants that disrupt binding of *UPF1*, a superfamily I RNA helicase, 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 and interacts with the exon junction complex[20]. While our finding seems unanticipated, it is consistent with recent work suggesting that *UPF1* binding to the 3'UTR can regulate target mRNA stability[21]. 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.

Variants that disrupt RBP binding influence risk of psychiatric disorders

Having established the importance of RBPs in selection and fitness, we next investigated the contribution of variants involved in RBP dysregulation to psychiatric disorder heritability. The high heritability estimates of psychiatric disorders make tracing major phenotypic outcomes of risk variants an important tool for understanding pathogenicity[22,23]. To address this question, we applied stratified LD score regression[24] as a statistical framework for partitioning 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^*, per-SNP heritability enrichment factor) standardized for comparison across different disease or trait-based 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)[25]. 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 studies[24–26]. Nevertheless, we performed a comprehensive negative control test in the context of RBP dysregulation. We simulated genetic architecture traits where the underlying casual SNPs were sampled entirely from experimentally profiled brain enhancers, promotors, and brain-expressed protein coding regions (i.e. mostly non-RBP regulatory regions) using real genotypes from the 1000 Genomes Project[27] (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: ADHD[28], autism spectrum disorder[29], bipolar disorder[30], major depression[31] and schizophrenia[32]. These GWAS were conducted with standardized analysis pipelines by the Psychiatric Genomics Consortium (PGC), minimizing 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 target site dysregulation for specific RBPs had a significant effect on psychiatric disorder risk after correcting for multiple hypothesis testing (Fig. 2a, 304 RBP-disease pairs with FDR < 0.05 after Benjamini-Hochberg correction, Supplementary Table 3,

Supplementary Fig. 3). These results indicate that risk variants for psychiatric disorders are extensively spread across RBP target regulatory networks, and biochemically underlie the polygenic architecture of mental disorders. In particular, we observed significantly larger per-SNP heritability effect sizes for RBPs with dysregulated target sites that are also differentially expressed in the developing human brain ($P < 2.2 \times 10^{-16}$, spearman rank test between τ^* and developmental target site overlap, Extended Data Fig. 2, Supplementary Fig. 4).

Additionally, we found novel associations between disrupted RBP target sites (*trans*-dysregulation) and RBPs that were themselves previously associated with disease (*cis*-dysregulation). For example, target site dysregulation of the spliceosome-associated factor *EFTUD2* contributed significantly across psychiatric disorders (*trans*-dysregulation, e.g. ADHD $p=1.4\times10^{-4}$, SCZ $p=6.4\times10^{-4}$ jackknife). Meanwhile exemplifying *cis*-dysregulation, haploinsufficiency of *EFTUD2* causes craniofacial malformation, microcephaly and developmental delay[9], phenotypes 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 brain[6] ($\tau^*=1.4$, $p=8.6\times10^{-3}$). Additionally, two *cis* GWAS loci within the *RBFOX1* RBP gene locus are associated with major depression risk[31]. Overall these data suggest that psychiatric disease risk is significantly linked with perturbations not only of RBPs (*cis*), but also the dysregulation of their targets (*trans*), which represent a much larger set of variants spread across the transcribed regions of the genome.

Post-transcriptional regulation broadly impacts psychiatric disorder risk

We find significant dysregulation effects across diverse biochemical regulatory categories of RBPs. For instance, for the most well-powered study, schizophrenia, we observed 49/91 significant RBP models (FDR < 0.05) that were UTR regulatory RBPs. This observation reveals the broad importance of post-transcriptional regulation beyond splicing, and was not limited to schizophrenia: the top psychiatric disorder-associated (TPA) RBPs (mean z-score > 2.5) covered RNA regulatory modes spanning splicing to transcript stability, based on the shared risk across all five psychiatric disorders studied (Fig. 2b). For example, *UPF1*[21] and *FAM120a*[33], which regulate transcript degradation, showed consistent, strong signals across these disorders, with top ranked effect sizes in schizophrenia (*UPF1* $\tau^*=1.16$ $p= 2.0\times10^{-6}$, *FAM120a* $\tau^*=1.2$ $p= 1.4\times10^{-7}$). We also observed a pair of ATP-dependent RNA helicases (*DDX6, DDX42*) among the top psychiatric disorder-associated RBPs. *De novo* mutations within *DDX6* have been shown to result in intellectual disability and developmental delay, adding support to the neuropathogenic role of this helicase[34]. Overall, these data demonstrate that disruption of diverse types of post-transcriptional regulation are highly associated with psychiatric disease risk.

Effects of RBP dysregulation exceed those of coding variants

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). Furthermore, within gene regions, the collective impact of disrupted RBP target sites can exceed coding region variant effects (Supplementary Fig. 5), thus providing further evidence that noncoding post-transcriptional regulation is a primary contributor to psychiatric disorder risk. 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).

RBP effects explain substantial heritability beyond known molecular QTLs

Previous reports have found that molecular QTLs are strongly enriched for disease heritability[13,35]. We therefore 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 GTEx[26], CommonMind[36] and BLUEPRINT[37] consortium (in addition to all baseline annotations). We found that the top psychiatric disorder-associated 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 thus provide an important additional tool for dissecting genetic architectures underlying disease.

Cross-ethnic replication

Finally, we sought to replicate our findings regarding the effects of RBP dysregulation using an independent cohort. We leveraged a recently published GWAS of East Asian schizophrenia participants[38] (22,778 schizophrenia cases and 35,362 controls), and compared our European PGC-schizophrenia RBP effect sizes with estimates that we obtained from the East Asian cohort. In this cross-ethnic replication analysis, we found highly significant concordance of RBP-associated risk between the two cohorts ($P < 2.2 \times 10^{-16}$ spearman rank test, see Extended Data Fig. 3 for RBP dysregulation effect sizes). Next, we tested if cross-disorder RBP effect size estimates from the Psychiatric Genomics Consortium cohort replicate in the iPSYCH cohort: a homogenous Danish population diagnosed using the same Danish public healthcare system criteria[39]. Likewise, we find consistent RBP dysregulation disease risk in the two independent cohorts ($P < 2.2 \times 10^{-16}$ spearman rank test, Extended Data Fig. 4). In summary, these replication analyses demonstrate that our RBP dysregulation disease risk estimates are concordant across populations with different genetic backgrounds.

RBP contributes to shared and distinct aspects of psychiatric disorders

In prior comparative analyses of genetic architectures, RBP biology has largely been ignored when stratifying genetic correlations into functional categories[22]. Here we sought to examine how variants disrupting RBP target sites shape shared and distinct genetic landscapes across psychiatric phenotypes. At the gene level, mutation-intolerant genes have been a shared source of enriched psychiatric disorder heritability[40]. We hypothesized that this enrichment could be in part driven by variants that affect RBP dysregulation. 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 ExAC[41], controlling for different baseline gene heritability enrichment levels Methods).

Beyond psychiatric disorders, we found that numerous RBP models were significantly associated with the heritability of a broad set of psychiatric and related phenotypes examined by the Brainstorm Consortium[26] (Extended Data Fig. 5, 856 RBP model-phenotype pairs FDR < 0.05 after Benjamini-Hochberg correction). In particular, psychiatric disorder-associated RBPs also showed the largest effect sizes for psychiatric phenotypes (Fig. 3b spearman correlation). This association was especially strong for cognitive-behavioral traits like “Cigarettes per day” (a common proxy for addictive risk-taking behavior[42]) and “Depressive symptoms” (a widely shared clinical feature for many psychiatric disorders[43]), whereas the non-brain related phenotypes displayed lower correlation between overall RBP effect sizes and psychiatric disorders (Fig. 3b). These results suggest that RBP dysregulation variants affect neuropathogenic pathways and are a significant driver of the high genetic correlations observed between disease and cognitive-behavioral traits (Extended Data Fig. 6 for disease-phenotype stratified genetic correlation).

Beyond the aforementioned similarities, risk odds correlation of RBP dysregulation variants also shows distinct clustering of psychiatric disorders (Extended Data Fig. 7). Thus, we tested whether distinct RBP target sites can help explain differences between psychiatric disorders. We found that RBP ILF3 target site dysregulation contributes to the differential liability between schizophrenia and bipolar disorder (Fig. 3c, two biological replicate ILF3 models highlighted, $p=4.1\times10^{-5}$ LoF intolerant genes, jackknife), extending recent findings by the Psychiatric Genomics Consortium study[30]. In addition to *trans*-dysregulation enrichment of ILF3 target sites, when inspecting *cis*-dysregulation effects, *ILF3* was the 5th most significantly associated gene locus in bipolar disorder, though it had no significant *cis*-association in the better-powered schizophrenia study[32] ($p=1.2\times10^{-9}$ MAGMA statistical framework[44]). Colocalization analysis with the GTEx cohort further supports this ILF3 *cis*-dysregulation differential association (extended results, Supplementary Fig. 6). As further independent evidence, the PsychENCODE cohort transcriptome-wide association study (TWAS) analysis[45] identified the *cis*-regulated expression of RBP ILF3 as significantly associated with risk of bipolar disorder, but no evidence of association was observed for schizophrenia (TWAS-bipolar ILF3 $p=3.9\times10^{-7}$, TWAS- schizophrenia ILF3 p=0.61, Fig. 3d). In conclusion, the molecular network composed of both RBP ILF3 and its *trans*-regulatory targets differentiates mechanisms underlying these two psychiatric disorders. Determining how the ILF3 network alters cellular functions in the brain can shed light on how genetics influences variations in clinical outcome.

Functional mapping identifies *DDHD2* as a schizophrenia risk gene

Hundreds of genomic regions are associated with the risk of psychiatric disorders, consistent with a polygenic architecture[32]. However, very few disease-associated regions have been mapped to their causal SNPs, and the underlying biochemical mechanism dissected. As a case study, we leveraged our ability to interrogate genome-wide, allele-specific RBP target site dysregulation to investigate a schizophrenia risk region.

The 8p12 genomic region was first identified as a significant schizophrenia risk region in the Han Chinese population[46] and was subsequently found to be significantly linked to this disorder in Europeans[40]. 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[47,48] (Fig. 4a, this top Seqweaver predicted SNP rs6981405 was a fine-mapped candidate SNP (95% credible set)[40]).

ENCODE[49] 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 robustly 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 that the A allele disrupts QKI binding. Furthermore, RNAi-mediated depletion of QKI led to elevated levels of DDHD2 mRNA in C allele genotype cells, but not in the homozygous A allele genotype line, where QKI binding is already disrupted (Fig. 4c,d). Thus, mutation of this SNP in DDHD2 mRNA disrupts QKI regulation.

DDHD2 is a principal brain triglyceride lipase, that when mutated causes a hereditary neurological disease, spastic paraplegia[50]. 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 implies 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 underlying molecular mechanisms. RBPs start acting on nascent RNA substrates, and influence every aspect in the life of a transcript, including protein expression and function. Therefore, interrogating genetic architectures at this early layer of molecular regulation is powerful because it reduces the complexity of identifying causal factors compared to further downstream approaches. Importantly, targeted biochemical perturbation of RBP-RNA interactions has a promising record in clinical intervention[51]. Therefore, establishing RBP dysregulation as a major source of molecular dysfunction contributing to psychiatric disorders and further identifying specific pathogenic target sites is a critical task.

In this work, we establish RBP dysregulation as a key factor affecting human fitness by identifying extensive negative selection signatures in the largest-to-date human WGS gnomAD[19] 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.

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 that emerges is the convergence of psychiatric disorder risk at both the RBP protein itself, and its target site. 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 molecular differences between schizophrenia and bipolar disorder. Similarly, the RBP "fragile X mental retardation protein" (FMRP) is the most common monogenic cause of autism[52], and FMRP mRNA targets are highly linked to both autism and schizophrenia[18,53,54]. Thus, these converging RBP regulatory networks may present ideal clinical targets, due to their greater collective biochemical contribution to pathogenicity.

Methodologically, we demonstrate that 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 widespread psychiatric disease risk is associated with RBPs that regulate processes across the life of the RNA. Current molecular QTL resources, while incredibly valuable, lack the breadth to capture these diverse molecular functions (e.g. we estimate for schizophrenia the collective contribution of RBP dysregulation exceeds that of GTEx sQTLs by over 4 times, Extended Data Fig. 8). This caveat limits the scope of analysis for disease, now encapsulating hundreds of thousands of cohort samples (e.g. UK biobank[55]). Indeed, increasing evidence supports an extensive pathogenic role for RBPs in diseases ranging from cancer[56], autoimmune disease[57,58] to myopathy[59]. Our computational framework enables study of RBP dysregulation in these and other disorders at a whole-genome-scale (e.g. significant RBP associations identified for non-brain phenotypes, Extended Data Fig. 5). 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 have 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, capturing 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 deep convolutional neural networks (CNN)[60] to build a quantitative model of RNA sequence features required for RBP binding, as biochemically assayed by CLIP[61] (training data). These RBP models subsequently enable probabilistic inference of the effect of sequence variants, capturing both direct and indirect effects, on RBP binding potential. The applied Seqweaver RBP model architecture and training are described in our previous *de novo* mutation autism work[16], and to ensure that our current results can be directly comparable, we used identical 232 CLIP-based RBP models (88 RBPs represented,

Supplementary Table 1) without any modifications (i.e. no additional training or manual parameter changes). All RBPs modeled in this study show robust expression in the brain and neuronal progenitor cells (Supplementary Fig. 7). RBP models based on mouse CLIP experiments were only included for RBPs that have shown to be conserved in humans[6], and replicate RBP models were treated separately by Seqweaver to provide users with maximum interpretability, linking them to specific CLIP experimental conditions.

Methodologically, CNNs allow researchers to design network architectures that can leverage information from high order motifs at different spatial scales but with optimal parameter sharing to avoid overfitting. Our Seqweaver CNN architecture consists of an initial input layer followed by a series of convolution and pooling layers. The input sequence layer contains a 4 × 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 an 8 bp window that shifts 1 bp at a time 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. ReLU(x) = max(0,x)). Thus formally, input S results in convolution layer output location n for kernel k as the following:

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

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 is collapsed into the maximum value observed in that span. The resulting output is then 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 and 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 are also 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. Importantly, 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 gnomAD[19] 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 described[16], annotating to 5'UTR, 3'UTR or 200bps introns flanking an exon previously observed to be alternatively spliced[62]. 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. This essentially involves querying for statistically significant interactions between a variant's gene location and the degree of selection acting against 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 effect sizes due to RBP dysregulation

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 for a disease, but may in reality be tagging enrichment of protein coding region SNPs due to high LD in the region. To resolve this challenge, we use the previously published statistical framework of stratified LD score regression[24] 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

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

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 bias[24]. 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 annotation[25].

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

More formally for RBP dysregulation annotations, $\boldsymbol{\tau}^*$ 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 the 1000 Genomes Project (European cohort), and fit the regression model only on HapMap SNPs with MAF > 0.05 as previously conducted[24]. The block jackknife procedure was used to test statistical deviation from zero for each fitted $\boldsymbol{\tau}^*$.

As presented in the regression model, we fit $\tau_{rbp}^*$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)[25], 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. 8. 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 conserved[63–65], nevertheless our conclusions are robust to the inclusion of conservation to baseline annotations (Extended Data Fig. 9). The final reported RBP effect size (τ_{rbp}^*) were obtained by jointly fitting, iteratively for each RBP, with all baseline annotations (full 71 baseline annotations listed in Supplementary Table 5, baseline annotation correlation with RBP annotations Supplementary Fig. 9).

Simulations for RBP effect size estimation

We conducted simulations to ensure that our regression models produced unbiased RBP effect sizes. Specifically, we verified 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 simGWAS[66]. 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 Consortium[67], and both promotors[68] and coding regions (restricted to nonsynonymous variants) that are expressed in the brain[69]. 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 were uniformly processed and analyzed, and were sufficiently powered to observe genome-wide significant SNPs. Datasets on five disorders met these criteria– 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 (excluding traits that did not find genome-wide significant SNPs). The East Asian schizophrenia cohort[38] and Danish cohort from the iPSYCH consortium cross-disorder GWAS study[39] was used for replication analysis. For cross-disorder replication comparison, Non-iPSYCH-overlapping PGC cohorts obtained from older PGC GWAS studies were used for cross-disorder RBP effect size estimates[32,70–73]. Population matched schizophrenia case vs. bipolar disorder case GWAS[30] 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 GTEx[26] eQTL (FE-meta) and BLUEPRINT[49] molecular QTL annotations were obtained from a previous study[35], that generated and validated the max causal posterior probability (MaxCPP)-based QTL annotations for GWAS enrichment analysis. The CommonMind[74] isoformQTLs and GTEx sQTL (brain cortex version 8) were fine-mapped to produce MaxCPP annotations following the same procedure as was previously reported[35]. 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.

RBP dysregulation effect comparison

For RBP effect size association with brain development, we obtained single bp-level genomic regions that are differentially differential expressed in the human prefrontal cortex brain with developmental stage and age from Jaffe et al[75]. Peak childhood stage annotations were obtained by filtering for regions that showed both differential expression and maximum expression during childhood in comparison to other age groups.

To estimate the collective RBP dysregulation effects in comparison to other annotations (e.g. sQTLs), we leveraged the out-of-sample East Asian schizophrenia GWAS[38]. To find the top combined RBP dysregulation effect SNPs associated with schizophrenia, we jointly fit all RBP models in the regression model to SNP chi-square values from the East Asian GWAS via LD score regression. Next we use the joint RBP regression model (i.e. weight each RBP model effects by the estimated regression coefficients) to find the top combined RBP dysregulation effect SNPs (top 0.1%, 0.5%, 1%) to examine the amount of heritability they capture compared to other annotations (e.g. sQTL) in the PGC European schizophrenia GWAS (i.e. non-overlapping cohort with the model fitting East Asian GWAS).

Genetic architecture analysis

Loss-of-function intolerant genes were obtained by the ExAC consortium with pLI threshold of > 0.9 as previously described[16]. 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.

MAGMA[44] was used to estimate the gene level association with schizophrenia and bipolar disorder. GENCODE[76] v25 gene annotations lifted to GRCh37 coordinates and total 19,984 protein coding genes were analyzed. We used SNPs from 10k upstream of the gene body and 1.5k downstream for each gene as previously used in a Psychiatric Genomics Consortium schizophrenia GWAS analysis study[77] along with the 1000 Genomes Project[27] European reference panel. Colocalization analysis for the *ILF3* locus between schizophrenia vs. bipolar GWAS[30] and GTEx ILF3 eQTL (v8 meta-tissue) was conducted using Coloc[78]. ILF3 TWAS results were obtained from the psychENCODE study[45].

Genetic correlation analysis stratified by RBP dysregulation was conducted by first fitting a joint RBP regression model for each disease/trait to estimate the collective phenotypic effect of RBP dysregulation for each SNP. Next when examining a pair of phenotypes (x,y), we iterate over each independent LD block[79] and extract the top RBP dysregulation variant that influence disorder x and their GWAS effect sizes on both x and y. We then calculated correlation between the GWAS effect sizes on x and the GWAS effect sizes on y.

Functional mapping of *DDHD2*

QKI eCLIP and knockdown RNA-seq data was obtained by the ENCODE project[49] in K562 and HepG2 cell lines. eCLIP data was processed as previously described[16] and visualized in IGV[80]. Kallisto[81] coupled with Sleuth[82] 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 method[83]. Genotyping results for SNP rs6981405 in K562 and HepG2 lines were obtained from ENCODE project.

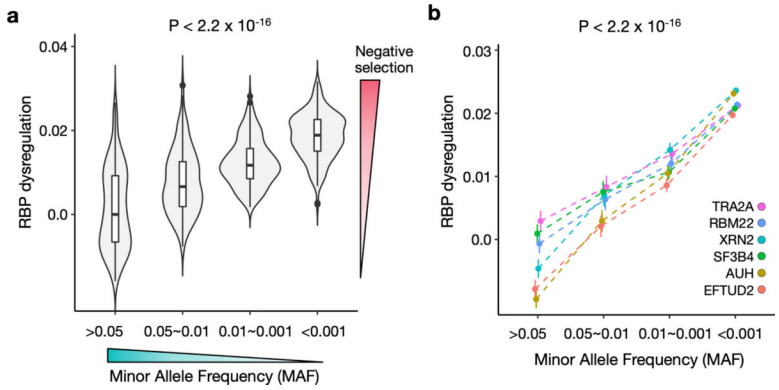
Data availability

All variant predicted scores have been made available to download and as an interactive web interface available at https://hb.flatironinstitute.org/seqweaver.

Code availability

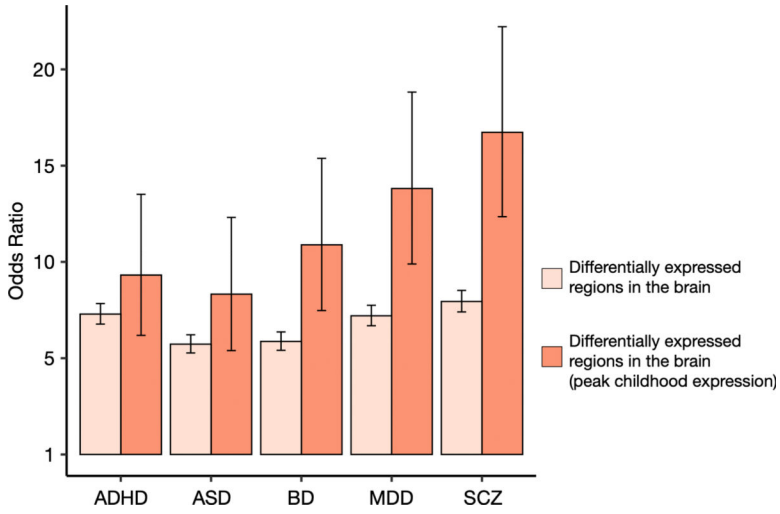
The code is available from https://hb.flatironinstitute.org/seqweaver/about.

Extended Data



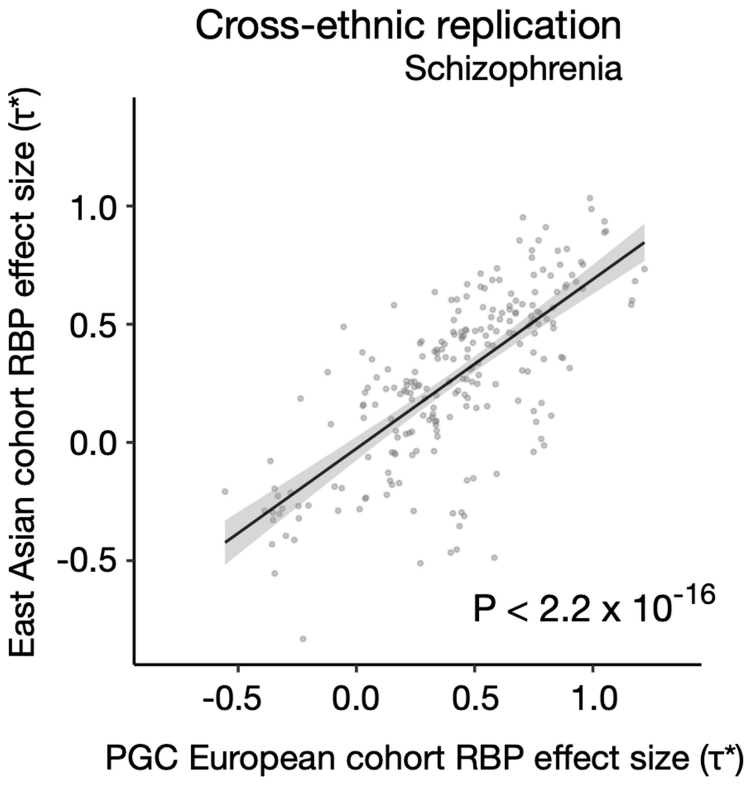
Extended Data Fig. 1. Population genetics revels negative selection acting on RBP target site dysregulation.

a) Across the Seqweaver profiled RBPs, we observe differential selection signatures for variants when segregated by their RBP target site dysregulation levels. Specifically, for gnomAD cohort noncoding variants (MAF bins x-axis), mean RBP dysregulation (Y-axis) shows an inverse relation with allele frequency, consistent with significant negative selection acting on high impact RBP disrupting variants. b) The top RBPs previously implicated by their autism *de novo* mutation risk (Zhou, Park, Theesfeld et al.), all show significant negative selection signatures, consistent with selection impeding RBP impacting variants from reaching high population prevalence. P-values from Wald test for slope and all inferred mean RBP dysregulation scores were normalized by subtracting average dysregulation predicted scores of common variants (MAF > 0.05) for comparison (95% CI).



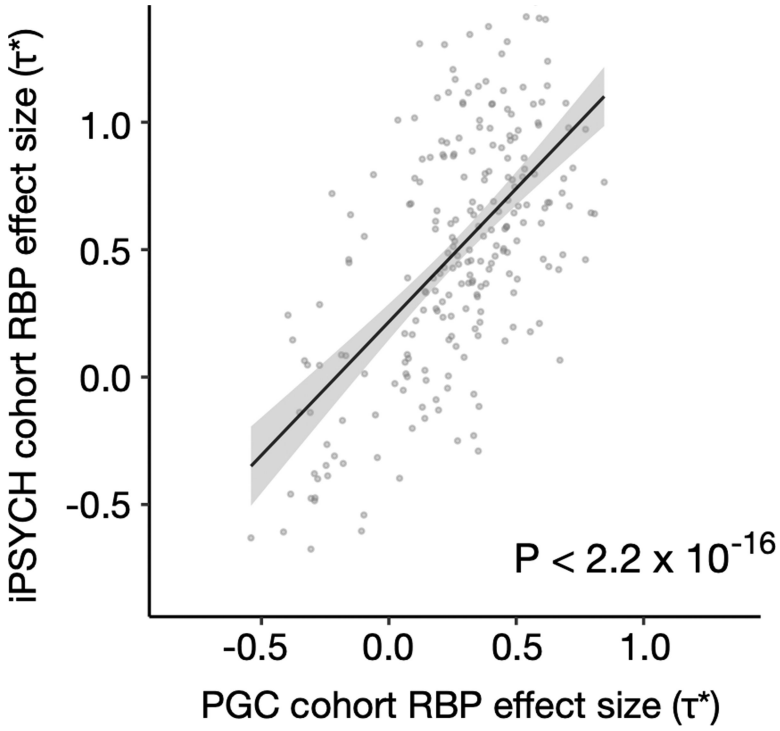
Extended Data Fig. 2. Regions with peak childhood stage expression shows the largest enrichment association with RBP dysregulation.

We test the overlap between prefrontal cortex brain differential expressed regions and RBP dysregulation SNPs (the top 0.5%) associated with each disorder in comparison to the genome-wide rate. We also plot the enrichment overlap for the subset of regions in which the expression was highest during childhood stage. All ORs have an enrichment p-value of $P < 2.2 \times 10^{-16}$. Error bars are 95% CI.



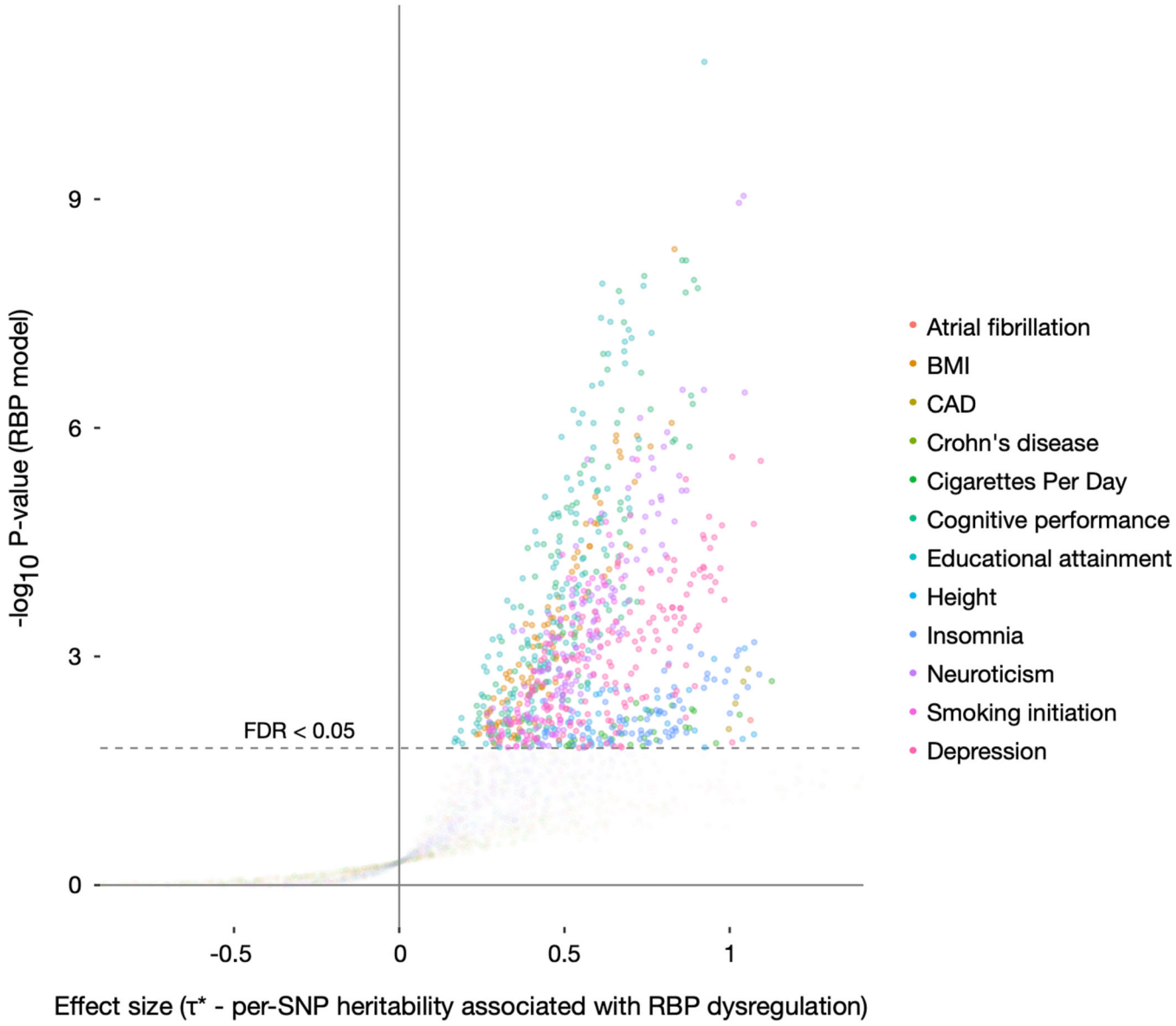
Extended Data Fig. 3. Cross-ethnic replication – RBP dysregulation effects replicate in an independent cohort.

Replication of estimated schizophrenia RBP dysregulation effect sizes (τ^*, European Psychiatric Genomics Consortium (PGC)) when compared to estimates from an East Asian cohort (Lam et al). P-value computed using spearman rank test of RBP effect sizes.

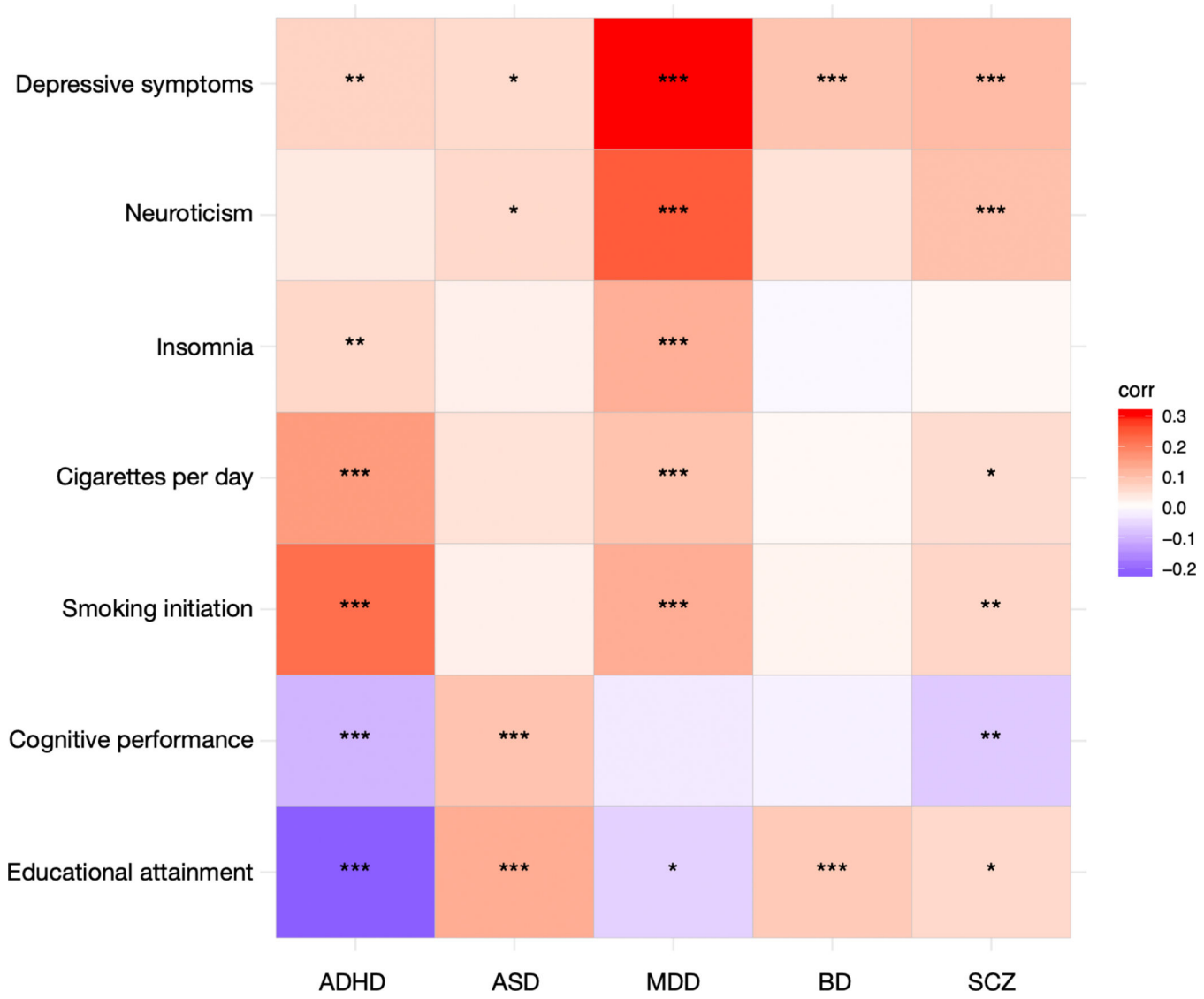


Extended Data Fig. 4. RBP dysregulation effects for cross-disorder risk replicate in iPSYCH cohort.

Replication of estimated cross-disorder RBP dysregulation effect sizes (τ^*, Psychiatric Genomics Consortium cohort) when compared to estimates from the iPSYCH cohort. P-value computed using spearman rank sum test of RBP effect sizes.

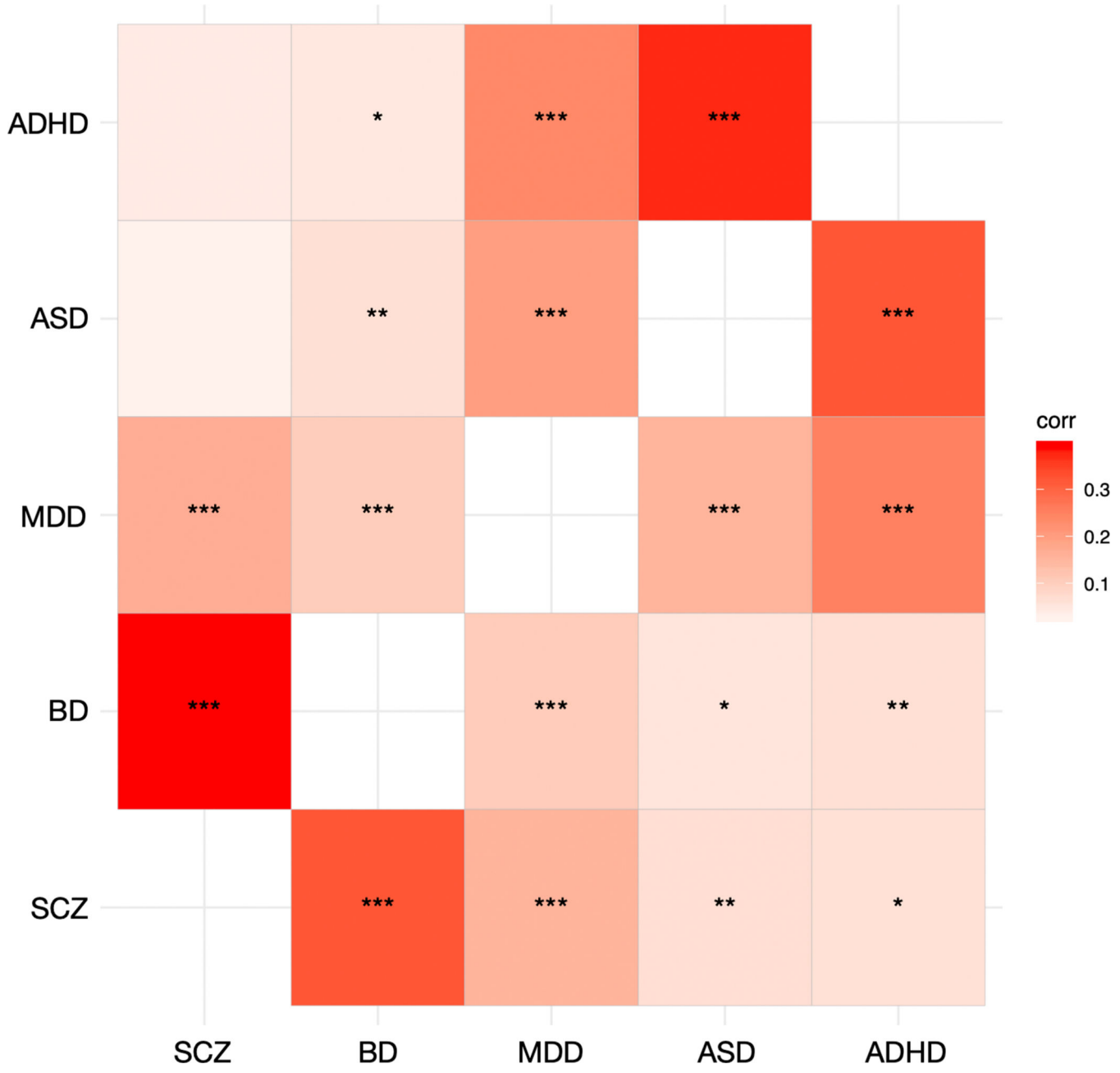


Extended Data Fig. 5. RBP dysregulation is a major contributor to human phenotypic variation.
The per-SNP heritability effect sizes (τ^*) for each RBP dysregulation is plotted across a collection of psychiatric traits, brain-associated anthropomorphic traits and representative non-brain related phenotypes previously examined by the Brainstorm Consortium study. 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).



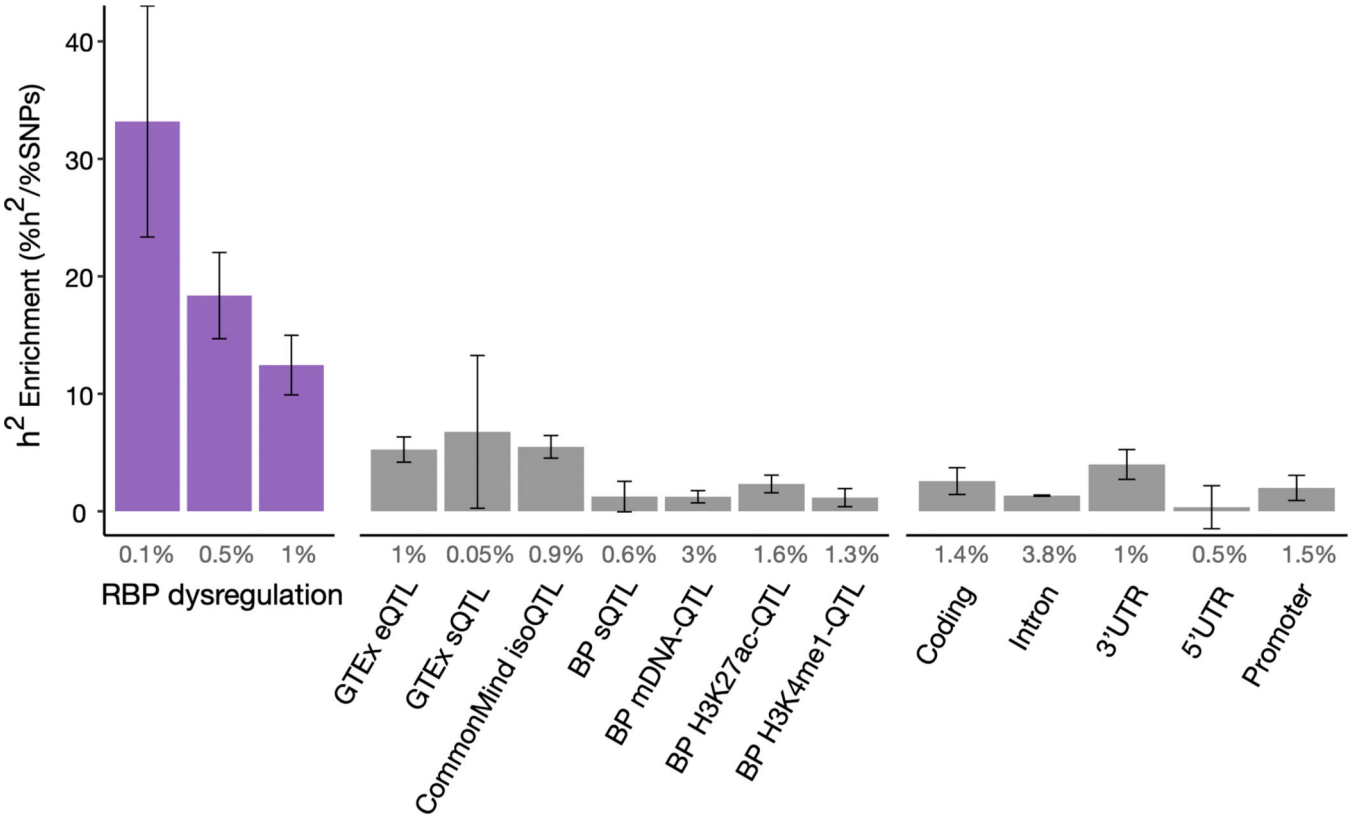
Extended Data Fig. 6. Heatmap showing patterns of correlated GWAS effect sizes between psychiatric disorders and behavioral-cognitive phenotypes for variants affecting RBP dysregulation.

For each pair of disorder and phenotype (x,y), we extracted the top RBP dysregulation set of variants that influence disorder x and their GWAS effect sizes on both x and y. We then calculated correlation between the GWAS effect sizes on x and the GWAS effect sizes on y, and tested whether this correlation was significantly different from zero. Stars represent statistical significance *** P< 0.001, ** P<0.01, * P<0.05.



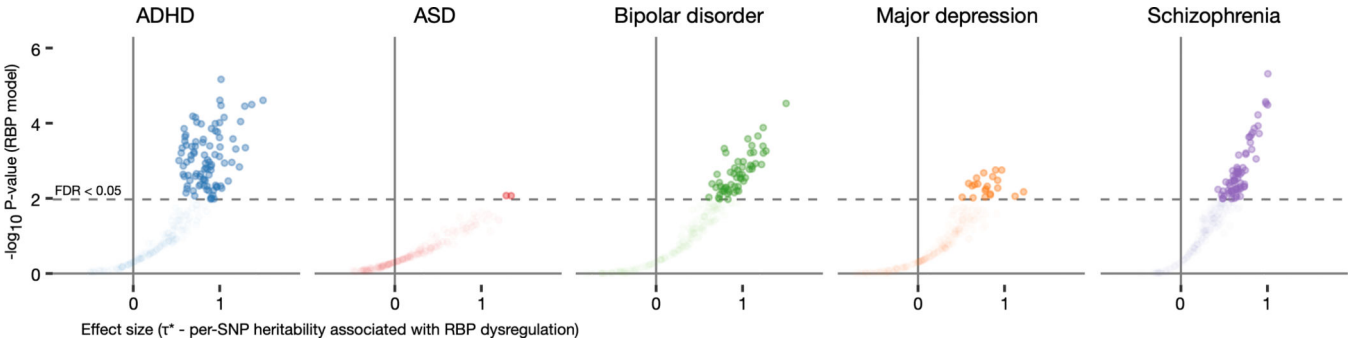
Extended Data Fig. 7. Heatmap showing patterns of correlated GWAS effect sizes between psychiatric disorders for variants affecting RBP dysregulation.

For each pair of disorders (x,y), we extracted the top RBP dysregulation set of variants that influence disorder x and their GWAS effect sizes on both x and y. We then calculated correlation between the GWAS effect sizes on x and the GWAS effect sizes on y, and tested whether this correlation was significantly different from zero. Stars represent statistical significance *** P< 0.001, ** P<0.01, * P<0.05.



Extended Data Fig. 8. Heritability enrichment for the collective RBP dysregulation effects in comparison to QTL and genomic functional annotations for schizophrenia.

The top 0.1%, 0.5%, 1% SNPs with the largest overall RBP dysregulation effects were compared to known molecular QTLs and gene/promoter annotations for their enrichment of heritability using PGC schizophrenia GWAS.



Extended Data Fig. 9. Estimated RBP dysregulation effects are robust after conditioning on conserved genomic elements.

The per-SNP heritability effect sizes (τ^*) for each RBP dysregulation is plotted across the five major psychiatric disorders after inclusion of vertebrate, mammal and primate conserved phastCons elements to the conditioning baseline annotation set (including QTL annotations). The dashed line indicates RBP models below FDR 0.05 threshold after multiple hypothesis correction (jackknife one-sided p-values; Benjamini-Hochberg correction).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Z. Zhang, A. Andersen and S. Lall for their help with the manuscript. We also thank all members of the Troyanskaya and Darnell laboratory for helpful discussions. This work is supported by NIH grants R01HG005998, U54HL117798 and R01GM071966, HHS grant HHSN272201000054C and Simons Foundation grant 395506. O.G.T. is a senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research (CIFAR). We thank the SFARI, Simons Foundation and Flatiron Institute. We are pleased to acknowledge that a substantial portion of the work in this paper was performed at the TIGRESS high-performance computer center at Princeton University, which is jointly supported by the Princeton Institute for Computational Science and Engineering and the Princeton University Office of Information Technology's Research Computing department.

References

1. Lee PH et al. Genome wide meta-analysis identifies genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. bioRxiv, 528117 (2019).
2. Black DL Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry 72, 291–336 (2003).
3. Krichevsky AM & Kosik KS Neuronal RNA Granules: A Link between RNA Localization and Stimulation-Dependent Translation. Neuron 32, 683–696 (2001). [PubMed: 11719208]
4. Guhaniyogi J & Brewer G. Regulation of mRNA stability in mammalian cells. Gene 265, 11–23 (2001). [PubMed: 11255003]
5. Costa-Mattioli M, Sossin WS, Klann E & Sonenberg N. Translational Control of Long-Lasting Synaptic Plasticity and Memory. Neuron 61, 10–26 (2009). [PubMed: 19146809]
6. Darnell RB RNA Protein Interaction in Neurons. Annual Review of Neuroscience 36, 243–270 (2013).
7. Schuman EM mRNA Trafficking and Local Protein Synthesis at the Synapse. Neuron 23, 645–648 (1999). [PubMed: 10482231]
8. Verkerk AJMH 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). [PubMed: 1710175]
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). [PubMed: 22305528]
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). [PubMed: 22541558]
11. Messiaen LM 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). [PubMed: 10862084]
12. Xiong HY et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806 (2015).
13. Li YI et al. RNA splicing is a primary link between genetic variation and disease. Science 352, 600 (2016). [PubMed: 27126046]
14. Jaganathan K et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535–548.e24 (2019).
15. Walker RL 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). [PubMed: 31133750]

17. Ule J et al. CLIP Identifies Nova-Regulated RNA Networks in the Brain. Science 302, 1212–1215 (2003). [PubMed: 14615540]
18. Iossifov I et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216–221 (2014). [PubMed: 25363768]
19. Karczewski KJ 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).
20. Kim YK & Maquat LE UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. RNA (New York, N.Y.) 25, 407–422 (2019).
21. Hogg JR & Goff SP Upf1 Senses 3′UTR Length to Potentiate mRNA Decay. Cell 143, 379–389 (2010). [PubMed: 21029861]
22. Anttila V et al. Analysis of shared heritability in common disorders of the brain. Science 360, eaap8757 (2018).
23. Laursen TM 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). [PubMed: 16061761]
24. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics 47, 1228–1235 (2015). [PubMed: 26414678]
25. Gazal S et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics 49, 1421–1427 (2017). [PubMed: 28892061]
26. Aguet F et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017). [PubMed: 29022597]
27. Auton A et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]
28. Demontis D et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature Genetics 51, 63–75 (2019). [PubMed: 30478444]
29. Grove J et al. Identification of common genetic risk variants for autism spectrum disorder. Nature Genetics 51, 431–444 (2019). [PubMed: 30804558]
30. Ruderfer DM et al. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. Cell 173, 1705–1715.e16 (2018).
31. Wray NR et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature Genetics 50, 668–681 (2018). [PubMed: 29700475]
32. Ripke S et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014). [PubMed: 25056061]
33. Kelly TJ, Suzuki HI, Zamudio JR, Suzuki M & Sharp PA Sequestration of microRNA-mediated target repression by the Ago2-associated RNA-binding protein FAM120A. RNA 25, 1291–1297 (2019). [PubMed: 31289130]
34. 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). [PubMed: 31422817]
35. 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). [PubMed: 29942083]
36. Fromer M et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience 19, 1442–1453 (2016). [PubMed: 27668389]
37. Chen L et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 167, 1398–1414.e24 (2016).
38. Lam M et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. Nature Genetics 51, 1670–1678 (2019). [PubMed: 31740837]
39. Schork AJ 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). [PubMed: 30692689]

40. Pardiñas AF et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics 50, 381–389 (2018). [PubMed: 29483656]
41. Lek M et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016). [PubMed: 27535533]
42. 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). [PubMed: 30643258]
43. Howard DM et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. Nature Communications 9, 1470 (2018).
44. de Leeuw CA, Mooij JM, Heskes T & Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Computational Biology 11, e1004219 (2015).
45. Gandal MJ et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science (New York, N.Y.) 362, eaat8127 (2018).
46. Shi Y et al. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. Nature genetics 43, 1224–1227 (2011). [PubMed: 22037555]
47. 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).
48. Bhalala OG, Nath AP, Consortium UKBE, Inouye M & Sibley CR Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. PLOS Genetics 14, e1007607 (2018).
49. Dunham I et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
50. Inloes JM 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).
51. Finkel RS et al. Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. New England Journal of Medicine 377, 1723–1732 (2017).
52. de la Torre-Ubieta L, Won H, Stein JL & Geschwind DH Advancing the understanding of autism disease mechanisms through genetics. Nature Medicine 22, 345–361 (2016).
53. Fromer M et al. De novo mutations in schizophrenia implicate synaptic networks. Nature 506, 179–184 (2014). [PubMed: 24463507]
54. Darnell Jennifer C. et al. FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. Cell 146, 247–261 (2011). [PubMed: 21784246]
55. 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).
56. 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). [PubMed: 24630730]
57. Darnell RB & Posner JB Paraneoplastic Syndromes Involving the Nervous System. New England Journal of Medicine 349, 1543–1554 (2003).
58. Häsler R et al. Alterations of pre-mRNA splicing in human inflammatory bowel disease. European Journal of Cell Biology 90, 603–611 (2011). [PubMed: 21324547]
59. Cummings BB et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science Translational Medicine 9, eaal5209 (2017).

Methods-only references

60. Krizhevsky A, Sutskever I & Hinton GE ImageNet classification with deep convolutional neural networks, 84–90 (Association for Computing Machinery, 2017).
61. Licatalosi DD et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456, 464–469 (2008). [PubMed: 18978773]
62. 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).

63. Sorek R & Ast G. Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse. Genome Research 13, 1631–1637 (2003). [PubMed: 12840041]
64. Lebedeva S et al. Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. Molecular Cell 43, 340–352 (2011). [PubMed: 21723171]
65. Lovci MT et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nature Structural & Molecular Biology 20, 1434–1442 (2013).
66. Fortune MD & Wallace C. simGWAS: a fast method for simulation of large scale case–control GWAS summary statistics. Bioinformatics 35, 1901–1906 (2018).
67. Wang D et al. Comprehensive functional genomic resource and integrative model for the human brain. Science 362, eaat8464 (2018).
68. Villar D et al. Enhancer Evolution across 20 Mammalian Species. Cell 160, 554–566 (2015). [PubMed: 25635462]
69. Kang HJ et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483–489 (2011). [PubMed: 22031440]
70. Anney RJL et al. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Molecular Autism 8, 21 (2017). [PubMed: 28540026]
71. Ripke S et al. A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry 18, 497–511 (2013). [PubMed: 22472876]
72. Neale BM et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 49, 884–97 (2010). [PubMed: 20732625]
73. Sklar P et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics 43, 977–983 (2011). [PubMed: 21926972]
74. Hoffman GE et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Scientific Data 6, 180 (2019). [PubMed: 31551426]
75. Jaffe AE et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nature Neuroscience 21, 1117–1125 (2018). [PubMed: 30050107]
76. Harrow J et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research 22, 1760–1774 (2012). [PubMed: 22955987]
77. Skene NG et al. Genetic identification of brain cell types underlying schizophrenia. Nature Genetics 50, 825–833 (2018). [PubMed: 29785013]
78. Giambartolomei C et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genetics 10, e1004383 (2014).
79. Berisa T & Pickrell JK Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32, 283–285 (2015). [PubMed: 26395773]
80. Thorvaldsdóttir H, Robinson JT & Mesirov JP Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178–192 (2012). [PubMed: 22517427]
81. Bray NL, Pimentel H, Melsted P & Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34, 525–527 (2016).
82. Pimentel H, Bray NL, Puente S, Melsted P & Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods 14, 687–690 (2017). [PubMed: 28581496]
83. Storey JD & Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100, 9440 (2003).

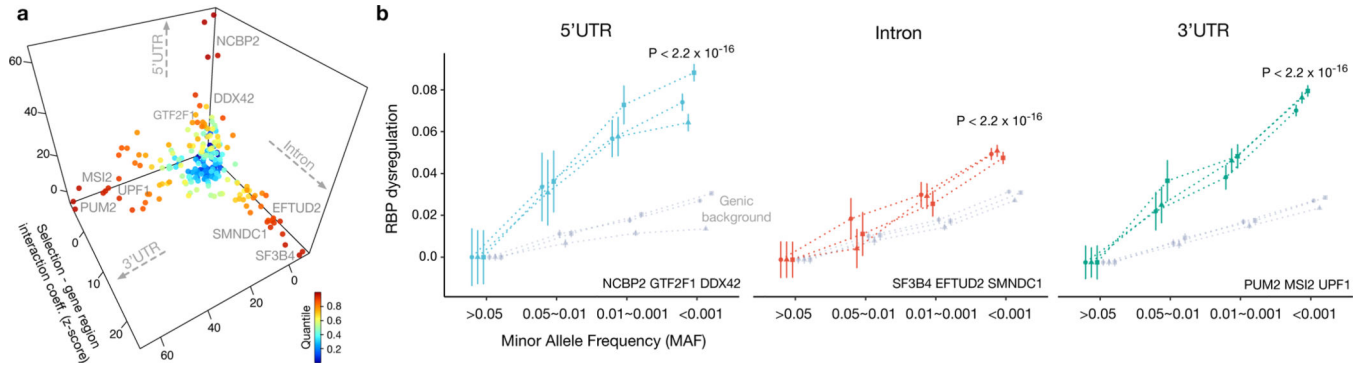


Figure 1. Negative selection signatures differentiate RBPs by their regulatory function.

a) For each RBP, the negative selection signatures are plotted along the major axes of sub-genic regulatory regions (5'/3' UTR and introns 200 bp flanking alternatively spliced exon). Higher coefficient z-scores, along the x, y or z axes, imply stronger regional selection compared to the baseline whole gene region (background). Fitness effects contributed by RBP functions beyond splicing are visualized by separation of RBPs along the regional selection axes. b) Examples of 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, dysregulation in z-score units) are shown. Stronger RBP selection signatures can be observed by the larger slope for each sub-genic region compared to the background gene levels (i.e. considering all variants in the gene). All inferred mean RBP dysregulation scores were normalized by subtracting average dysregulation predicted scores of common variants (>0.05 MAF) for comparison (95% CI).

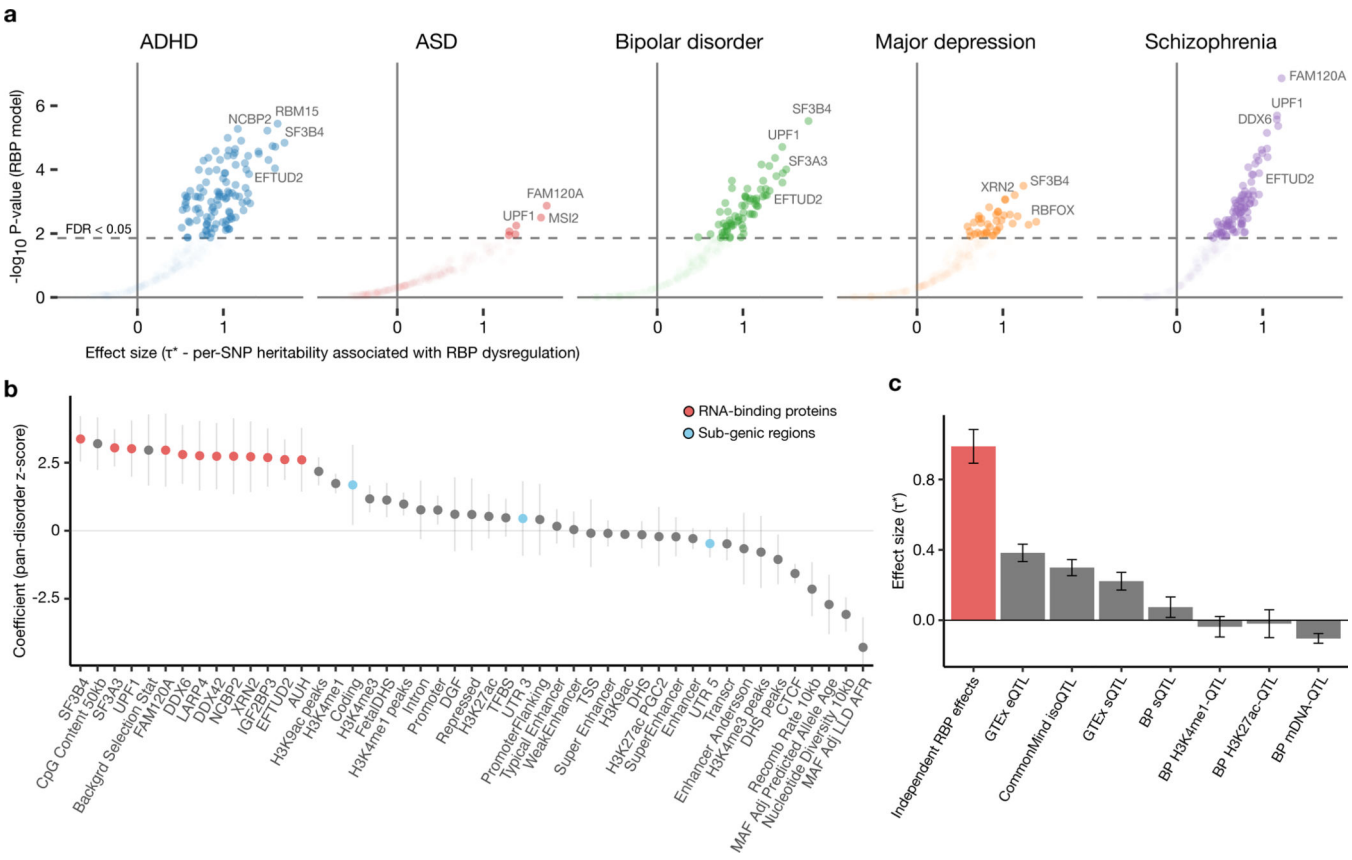


Figure 2. Genome-wide RBP dysregulation is a significant source of psychiatric disorder heritability.

a) The per-SNP heritability effect sizes (τ^*) for the target site dysregulation of each RBP 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). Negative effect size (τ^*) estimates represent the depletion of heritability for an annotation. 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). Annotation coefficients are obtained from LD score regression fit. 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.

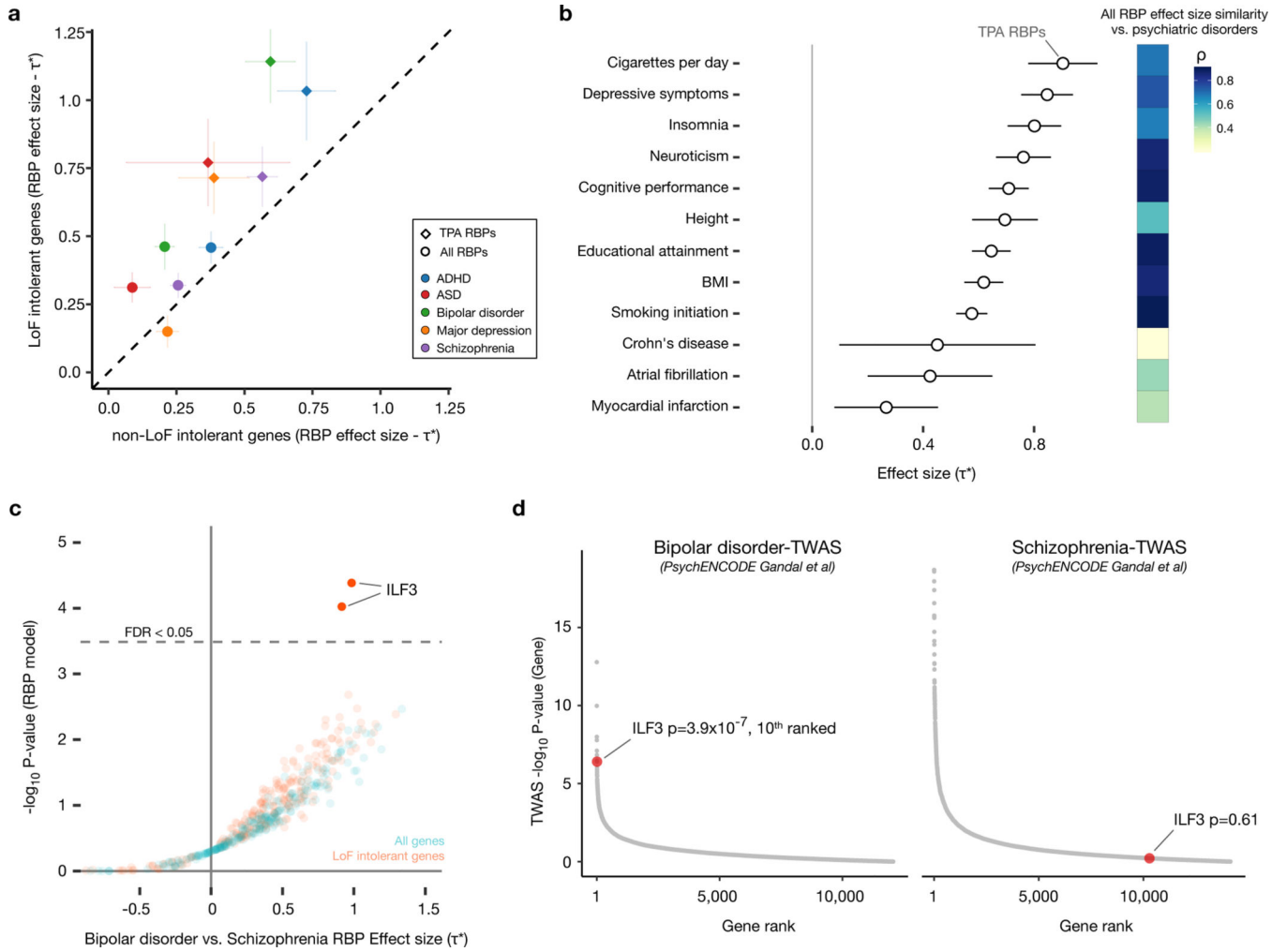


Figure 3. RBP dysregulation underlies shared and distinct genetic architectures of psychiatric phenotypes.

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 both for all genes and for LoF intolerant genes. The dashed line indicates RBP models below FDR 0.05 threshold (Benjamini-Hochberg method). d) TWAS analysis results from the PyschENCODE (Gandal et al) is shown by ranked gene-level statistical association for bipolar and schizophrenia (ILF3 highlighted).

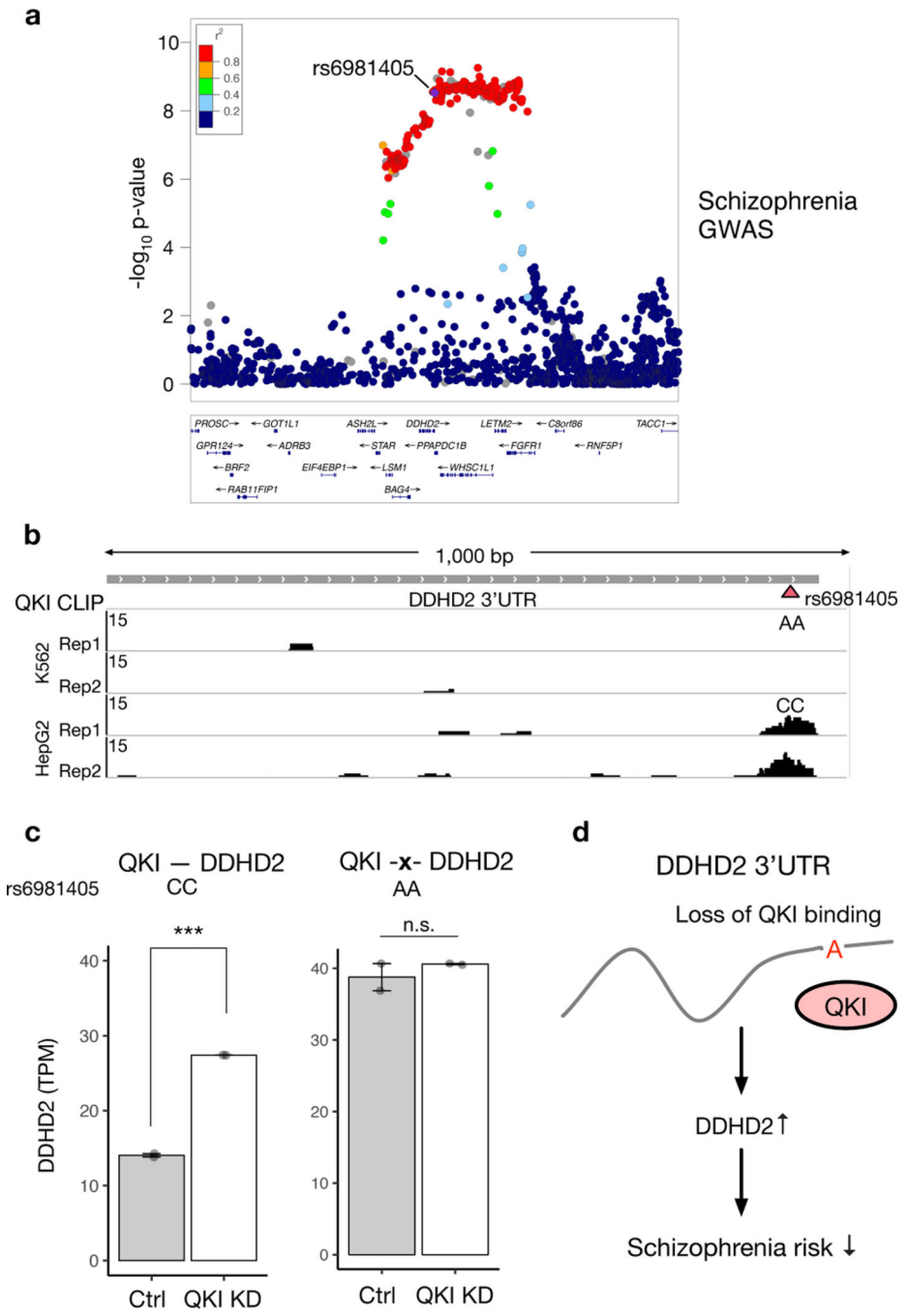


Figure 4. Functional RBP regulatory mapping identifies a schizophrenia risk variant in *DDHD2* 3'UTR.

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 followed by RNA-seq confirms 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.