



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

Christopher Y. Park¹✉, Jian Zhou¹, Aaron K. Wong¹, Kathleen M. Chen¹, Chandra L. Theesfeld², Robert B. Darnell^{3,4} and Olga G. Troyanskaya^{1,2,5}✉

Despite the strong genetic basis of psychiatric disorders, the underlying molecular mechanisms are largely unmapped. RNA-binding proteins (RBPs) are responsible for most post-transcriptional regulation, from splicing to translation 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 the 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 quantitative trait loci studies and has a stronger impact than common coding region variants. We share the genome-wide profiles of RBP dysregulation, which we use to identify *DDHD2* as a candidate schizophrenia risk gene. This resource provides a new analytical framework to connect the full range of RNA regulation to complex disease.

Interrogating the genetics underlying of psychiatric disorders is a key path to understanding the pathophysiological cause of mental illness. In particular, genome-wide association studies (GWAS) have become a widely adopted approach for studying the genetics of human psychiatric disorders. With ever-increasing sample cohorts, approaching one million individuals for psychiatric disorders¹, numerous risk loci have now been cataloged. Despite this progress, the biochemical perturbations and pathological mechanisms underlying human psychiatric disorders are challenging to decipher, critical as this is for translating genetic discoveries into sorely needed actionable targets.

RBPs regulate various aspects of RNA metabolism, including RNA splicing², localization³, stability⁴ and translation⁵. Each of these functions are critical not only for protein expression but also for proper spatiotemporal function, especially in the brain⁶. 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 nucleus⁷.

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 is 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 (that is, 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 signals in autism in probands versus their unaffected siblings¹⁶. In this study, 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, which are publicly available at hb.flatironinstitute.org/seqweaver. We show that across RBPs with diverse functions, dysregulated RBP target sites are top drivers of psychiatric disorder risk. Leveraging this resource, we discover a new link between an RBP-disrupting variant in *DDHD2*, encoding a phospholipase involved in hereditary spastic paraplegia, and a multi-ancestry-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, Seqweaver¹⁶, that can accurately estimate RBP binding based on sequence information (Supplementary Fig. 1), trained on *in vivo* RBP–RNA interaction profiles (cross-linking immunoprecipitation (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); thus,

¹Flatiron Institute, Simons Foundation, New York, NY, USA. ²Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA.

³Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, NY, USA. ⁴Howard Hughes Medical Institute, Chevy Chase, MD, USA.

⁵Department of Computer Science, Princeton University, Princeton, NJ, USA. ✉e-mail: cpark@flatironinstitute.org; ogt@cs.princeton.edu

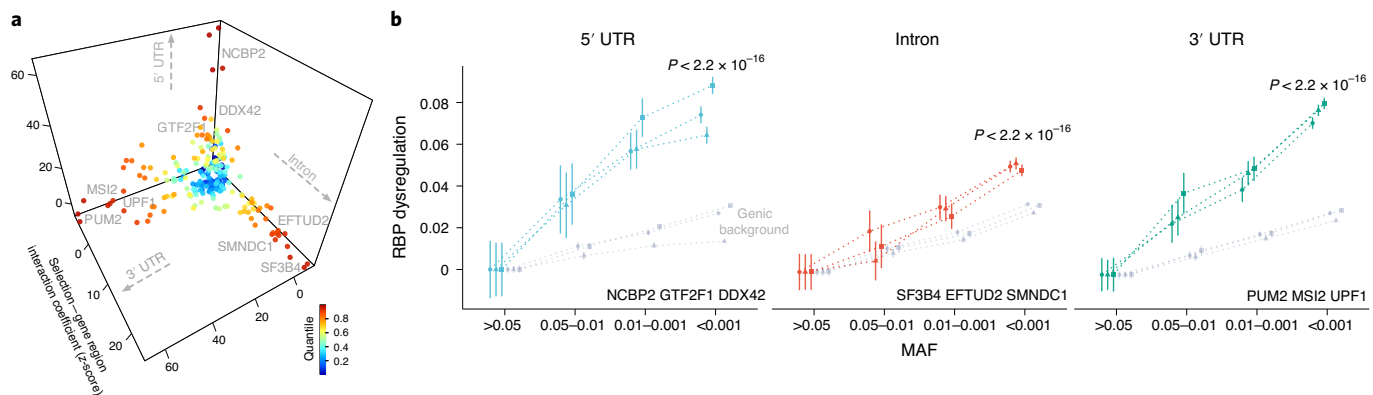


Fig. 1 | Negative selection signatures differentiate RBPs by their regulatory function. a, For each RBP, the negative selection signatures were plotted along the major axes of subgenic regulatory regions (5'/3' UTR and introns 200 bp flanking the alternatively spliced exon). Higher coefficient z-scores along the x, y or z axes imply stronger regional selection compared to the baseline whole-gene region (background). The fitness effects contributed by RBP functions beyond splicing are visualized by separation of RBPs along the regional selection axes. **b**, Examples of subgenic 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 subgenic region compared to the background gene levels (that is, considering all variants in the gene). All inferred mean RBP dysregulation scores were normalized by subtracting the average dysregulation predicted scores of common variants (MAF > 0.05) for comparison (95% confidence interval (CI)).

individual RBP phenotypic effects can be mapped to these genetic changes. Because our models enable the 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 RBP regulation. Case studies exploring pathogenic de novo mutations¹⁸ predicted 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 (the RBP model list is shown in Supplementary Table 1). We found significant depletion of strong effect RBP dysregulation variants at high allele frequencies (minor allele frequency (MAF) > 0.05, $P < 2.2 \times 10^{-16}$ Wilcoxon rank-sum test common versus ultrarare (MAF < 0.001)). The mean impact of target site variants increased significantly from common (MAF > 0.05) to ultrarare 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 messenger RNA 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 subgenic 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 false discovery rate (FDR) < 0.05

subgenic annotation interaction; full results in Supplementary Table 2). The significant association of each RBP with a subgenic location rules out the baseline interpretation of confounding stochastic nonfunctional interactions and points to an active regulatory function with specific fitness consequences.

For many known splicing factors (for example, 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 MAFs 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 poly(A) regulatory proteins (for example, MSI2, PUM2, PABP and ELAVL1; all $P < 2.2 \times 10^{-16}$).

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 complex²⁰. While our finding seems unanticipated, it is consistent with recent work suggesting that UPF1 binding to the 3' UTR can regulate target mRNA stability²¹. 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 the 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 linkage disequilibrium (LD) score regression²⁴ 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 the estimation of SNP effects (τ , per-SNP heritability enrichment factor) standardized for comparison across different disease

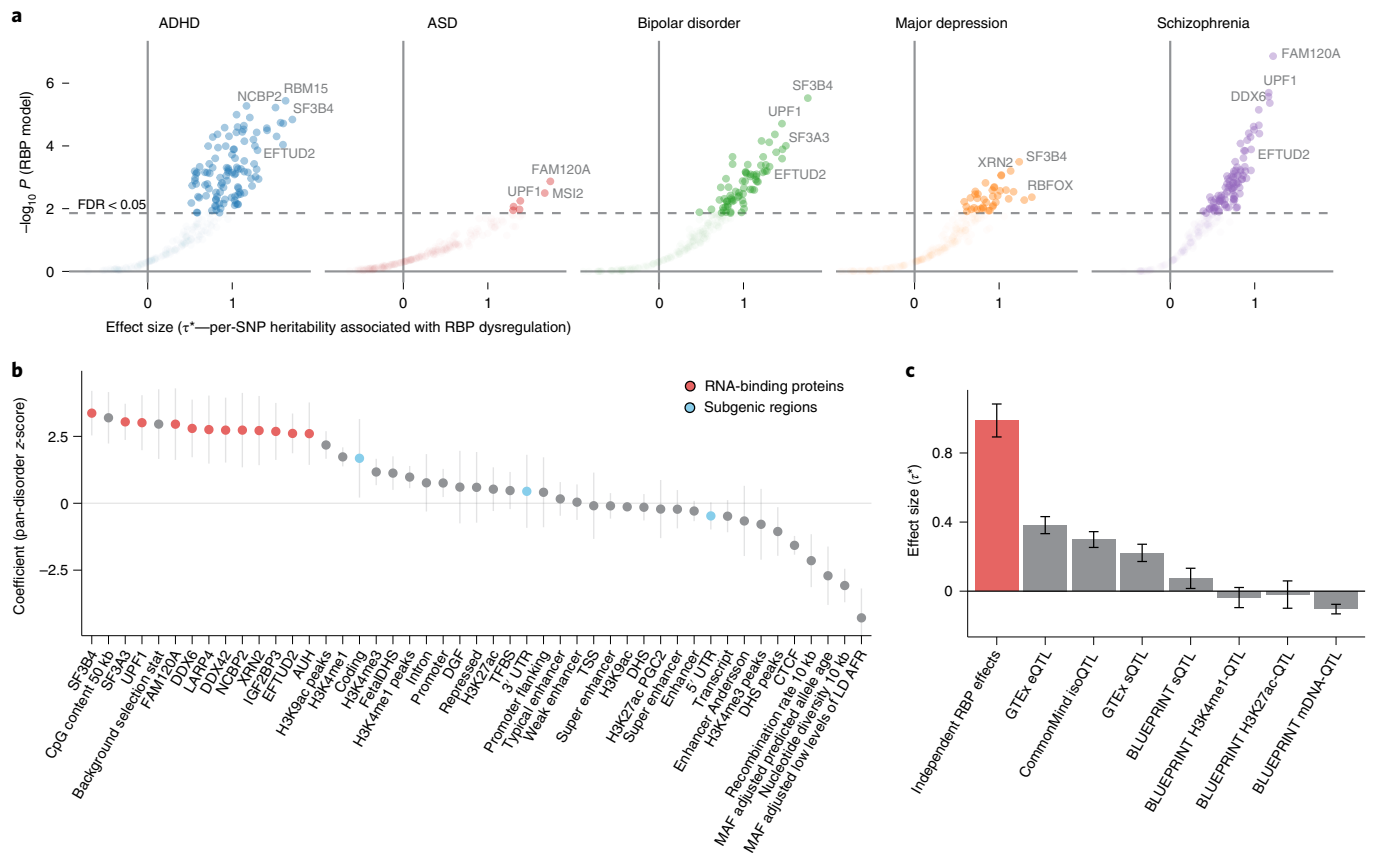


Fig. 2 | Genome-wide RBP dysregulation is a significant source of psychiatric disorder heritability. **a**, The per-SNP heritability effect sizes (τ^*) for the target site dysregulation of each RBP are plotted across the five major psychiatric disorders. The dashed line indicates RBP models below an $FDR < 0.05$ 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 RBPs (mean z-scores across the five disorders). The jointly fitted representative baseline annotations are also shown for comparison (subgenic region annotations are highlighted in blue, for example, coding region). Annotation coefficients were obtained from the LD score regression fit. **c**, Per-SNP heritability effect sizes (τ^*) for top psychiatric-disorder-associated RBPs after conditioning on a collection of molecular QTL annotations (that is, independent RBP effects from molecular QTLs and baseline annotations). The jointly fitted collection of QTL annotation effect sizes is also plotted. mRNA isoform (isoQTL), DNA methylation (mDNA QTL). All error bars are 95% CI.

or trait-based GWAS studies while conditioning on a collection of baseline functional annotations (for example, coding region, allele age, CpG content, enhancers, promoter and epigenetic histone marks; Methods)²⁵. In this study, we combined 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 causal SNPs were sampled entirely from experimentally profiled brain enhancers, promoters and brain-expressed protein-coding regions (that is, mostly non-RBP regulatory regions) using real genotypes from the 1000 Genomes Project²⁷ (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: attention deficit hyperactivity disorder (ADHD)²⁸; autism spectrum disorder (ASD)²⁹; bipolar disorder³⁰; major depression³¹; and schizophrenia³². 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 5 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 an $FDR < 0.05$ after Benjamini–Hochberg correction; Supplementary Table 3 and 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 and Supplementary Fig. 4).

Additionally, we found new 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, for example, ADHD $P = 1.4 \times 10^{-4}$,

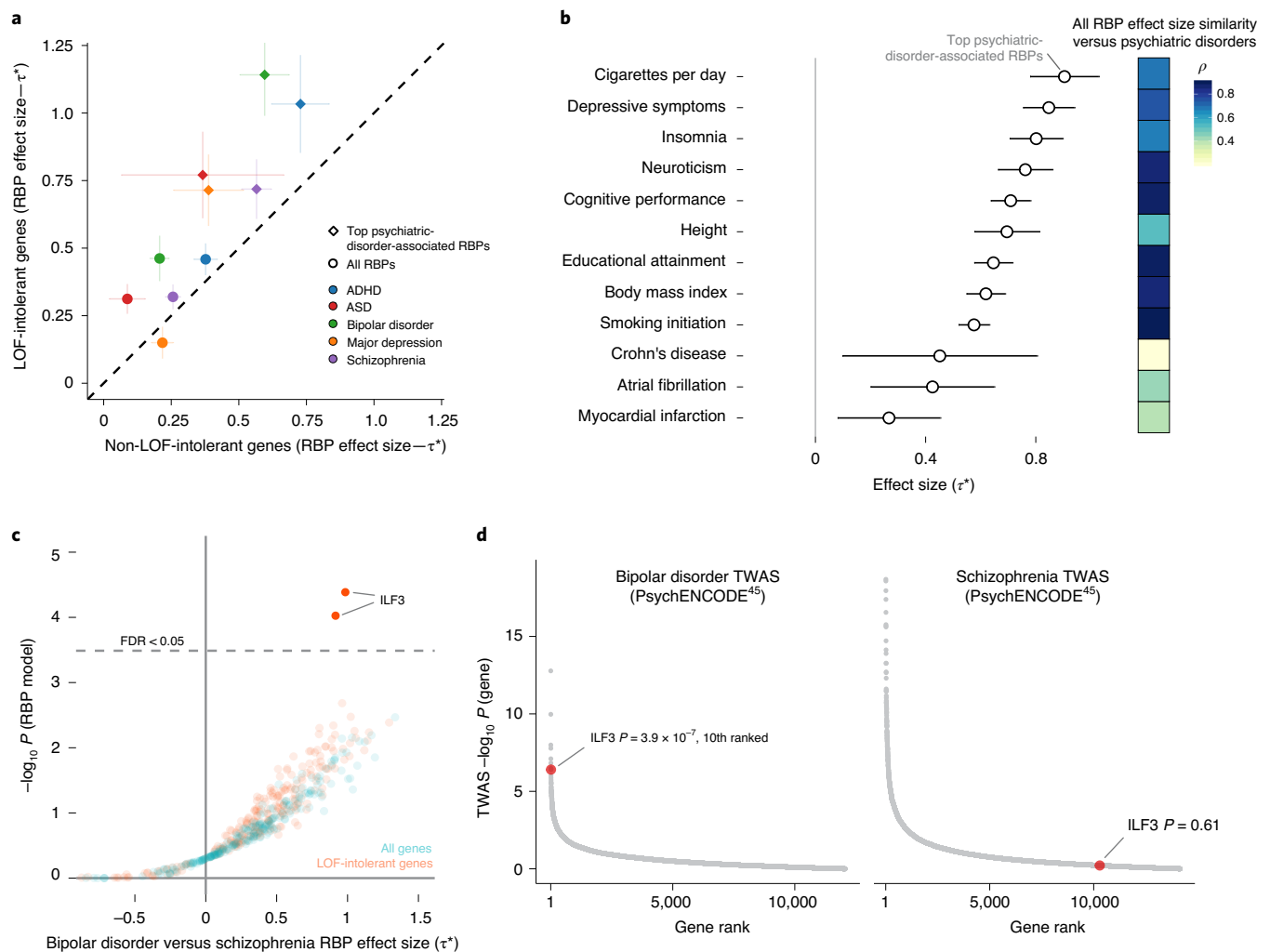


Fig. 3 | RBP dysregulation underlies shared and distinct genetic architectures of psychiatric phenotypes. **a**, The meta-analysis of RBP dysregulation effect sizes in LOF-intolerant genes compared to the remaining set of genes is shown (95% CI). All effect sizes are conditioned on baseline annotations plus two additional annotations delineating LOF-intolerant genes and their coding regions to control for the higher background heritability rates. **b**, Top psychiatric-disorder-associated 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 an FDR threshold of <0.05 (Benjamini-Hochberg method). **d**, TWAS analysis results from the PsychENCODE⁴⁵ is shown by ranked gene-level statistical association for bipolar disorder and schizophrenia (ILF3 is highlighted).

schizophrenia $P = 6.4 \times 10^{-4}$ jackknife). Meanwhile, exemplifying *cis*-dysregulation, haploinsufficiency of *EFTUD2* causes craniofacial malformation, microcephaly and developmental delay⁹, 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 brain⁶ ($\tau^* = 1.4$, $P = 8.6 \times 10^{-3}$). Additionally, two *cis* GWAS loci within the *RBFOX1* RBP gene locus were associated with major depression risk³¹. 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 found significant dysregulation effects across diverse biochemical regulatory categories of RBPs. For instance,

for the most well-powered study, schizophrenia, we observed 49 out of 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 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 (ref. ²¹) and FAM120A³³, which regulate transcript degradation, showed consistent, strong signals across these disorders, with top ranked effect sizes in schizophrenia (*UPF1* $\tau^* = 1.16$ $P = 2.0 \times 10^{-6}$; *FAM120A* $\tau^* = 1.2$ $P = 1.4 \times 10^{-7}$). We also observed a pair of ATP-dependent RNA helicases (*DDX6*, *DDX42*) among the 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 helicase³⁴. Overall, these data demonstrate that disruption of diverse

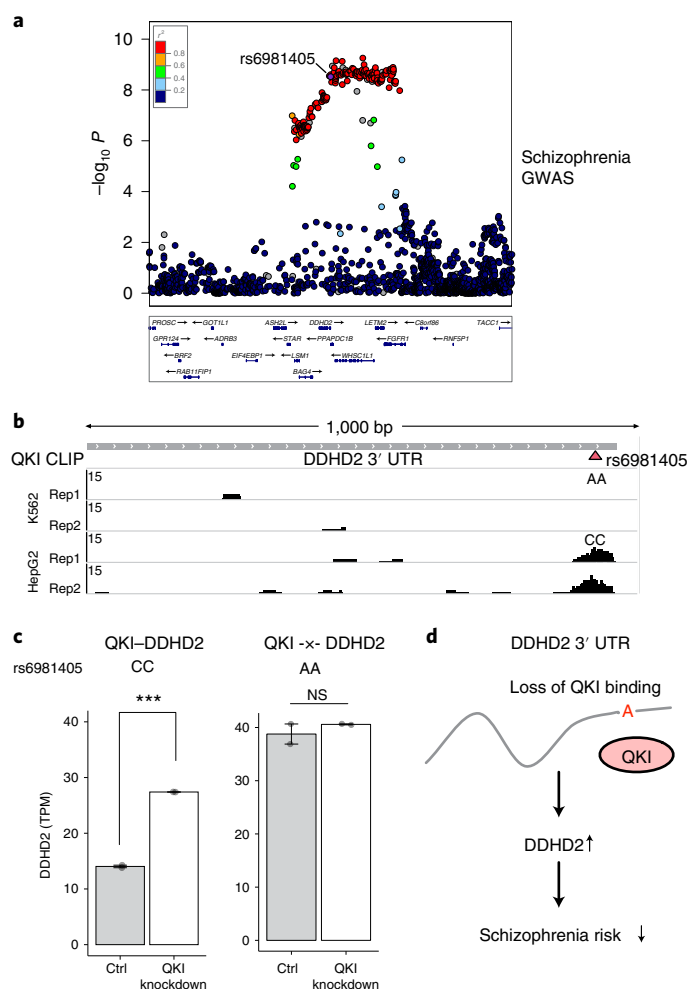


Fig. 4 | Functional RBP regulatory mapping identifies a schizophrenia risk variant in the *DDHD2* 3' UTR. **a**, Schizophrenia GWAS signal for the cross-ancestry-associated region. The highlighted SNP **rs6981405** represents the top predicted RBP dysregulation variant disrupting RBP QKI binding. **b**, ENCODE eCLIP data confirming that SNP **rs6981405** C>A leads to the disruption of RBP QKI binding to *DDHD2* (**rs6981405** genotype for cell lines K562 homozygous AA and HepG2 homozygous CC). SNPs were allowed during CLIP read alignment. **c**, QKI knockdown followed by RNA sequencing confirms QKI-mediated regulation of *DDHD2*, which is disrupted in the homozygous AA genotype (that is, QKI knockdown shows no effect when SNP **rs6981405** impedes with RBP binding). Homozygous CC $P = 0.001$, q value = 0.014. The error bars represent the s.e.m. **d**, The variant at SNP **rs6981405** disrupts the QKI-*DDHD2* 3' UTR interaction, which alters the abundance of mature *DDHD2* mRNA, and, in turn, schizophrenia risk.

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 fitted annotations in the regression models. We found that the statistical association between disease heritability and RBPs is among the top functional annotation, 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 the 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 are significant and robust after conditioning on potential confounding factors, such as background selection rate, low levels of LD, allele age and MAF (Fig. 2b; baseline annotations included in the regression model; Methods).

RBP effects explain substantial heritability beyond known molecular quantitative trait loci. Previous reports have found that molecular quantitative trait loci (QTLs) are strongly enriched for disease heritability^{13,35}. Therefore, 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 the Genotype-Tissue Expression (GTEx) project³⁶, CommonMind³⁶ and BLUEPRINT³⁷ consortia (in addition to all baseline annotations). We found that the top psychiatric-disorder-associated RBPs were highly significant and displayed overall greater effect sizes compared to the QTL annotations (Fig. 2c and Supplementary Table 4). Importantly, this suggests 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-ancestry 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 participants with schizophrenia³⁸ (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-ancestry 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 the RBP dysregulation effect sizes). Next, we tested whether cross-disorder RBP effect size estimates from the PGC cohort would be replicated in the iPSYCH cohort, a homogenous Danish population diagnosed using the same Danish public health-care system criteria³⁹. Likewise, we found 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 previous comparative analyses of genetic architectures, RBP biology has largely been ignored when stratifying genetic correlations into functional categories²². In this study, 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 heritability⁴⁰. 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 ExAC⁴¹, 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 Consortium²⁶ (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, such as ‘cigarettes per day’ (a common proxy for addictive risk-taking behavior⁴²) and ‘depressive symptoms’ (a widely shared clinical feature for many psychiatric disorders⁴³), whereas 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 showed 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 interleukin enhancer-binding factor 3 (ILF3) target site dysregulation contributes to the differential liability between schizophrenia and bipolar disorder (Fig. 3c; two biological replicate ILF3 models highlighted, $P = 4.1 \times 10^{-5}$ LOF-intolerant genes, jackknife), extending recent findings by the PGC study³⁰. 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, although it had no significant *cis*-association in the better-powered schizophrenia study³² ($P = 1.2 \times 10^{-9}$ MAGMA statistical framework⁴⁴). Colocalization analysis with the GTEx cohort further supported this ILF3 *cis*-dysregulation differential association (extended results in Supplementary Fig. 6). As further independent evidence, the PsychENCODE cohort transcriptome-wide association study (TWAS) analysis⁴⁵ identified the *cis*-regulated expression of RBP ILF3 as significantly associated with the risk of bipolar disorder; however, no evidence of association was observed for schizophrenia (bipolar disorder TWAS ILF3, $P = 3.9 \times 10^{-7}$, schizophrenia TWAS ILF3 $P = 0.61$; Fig. 3d). In conclusion, the molecular network composed of both RBP ILF3 and its *trans*-regulatory targets differentiates the 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, which is consistent with a polygenic architecture³². However, very few disease-associated regions have been mapped to their causal SNPs and the underlying biochemical mechanism dissected. In 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 population⁴⁶ and was subsequently found to be significantly linked to this disorder in Europeans⁴⁰. Cross-population replication suggests a robust molecular cause underlying the associated loci with global clinical potential. Our analysis provides a potential biochemical mechanism for this association. Within this region we identified an SNP in the *DDHD2* 3' UTR that can disrupt binding by the RBP protein quaking (QKI), which is known to play an important role in schizophrenia^{47,48} (Fig. 4a; this top Seqweaver predicted that SNP rs6981405 was a fine-mapped candidate SNP (95% credible set)⁴⁰).

ENCODE⁴⁹ QKI eCLIP data, available for 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 transcripts per million (TPM)). Importantly, QKI–*DDHD2* binding is observed only in the homozygous C allele genotype (Fig. 4b), which is consistent with our estimation that the A allele disrupts QKI binding. Furthermore, RNA-interference-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 paraplegia⁵⁰. Our genetic evidence, provided by RBP regulatory mapping, coupled with supporting experimental data, suggests 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 intervention⁵¹. 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 established RBP dysregulation as a key factor affecting human fitness by identifying extensive negative selection signatures in the largest-to-date human whole-genome sequencing gnomAD¹⁹ cohort. We further found that the 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 an extensive 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 found converging evidence linking RBP ILF3 and its RNA targets to molecular differences between schizophrenia and bipolar disorder. Similarly, the RBP ‘fragile X mental retardation 1’ (FMRP) protein is the most common monogenic cause of autism⁵² 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 demonstrated that deep learning inference of genome-wide molecular effects allows us to estimate major modes of biochemical perturbation and their contribution to disease. We found that splicing disruption is the tip of the iceberg since 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 (for example, we estimated for schizophrenia that the collective contribution of RBP dysregulation exceeds that of GTEx splicing QTLs (sQTLs) by over four times; Extended Data Fig. 8). This caveat limits the scope of analysis for disease, now encapsulating hundreds of thousands of cohort samples (for example, the UK Biobank⁵⁵). Indeed, increasing evidence supports an extensive pathogenic role for RBPs in diseases ranging from cancer⁵⁶, autoimmune disease^{57,58} to myopathy⁵⁹. Our

computational framework enables the study of RBP dysregulation in these and other disorders at a whole-genome scale (for example, significant RBP associations identified for non-brain phenotypes; Extended Data Fig. 5). Furthermore, as more tissue- and cell type-specific CLIP data become 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 ultrarare variants, should provide the means to interrogate RBP-derived human diseases at an unprecedented scale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-020-00761-3>.

Received: 19 February 2020; Accepted: 8 December 2020;
Published online: 18 January 2021

References

- Lee, P. H. et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482.e11 (2019).
- Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
- Krichevsky, A. M. & Kosik, K. S. Neuronal RNA granules: a link between RNA localization and stimulation-dependent translation. *Neuron* **32**, 683–696 (2001).
- Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11–23 (2001).
- Costa-Mattioli, M., Sossin, W. S., Klann, E. & Sonenberg, N. Translational control of long-lasting synaptic plasticity and memory. *Neuron* **61**, 10–26 (2009).
- Darnell, R. B. RNA protein interaction in neurons. *Annu. Rev. Neurosci.* **36**, 243–270 (2013).
- Schuman, E. M. mRNA trafficking and local protein synthesis at the synapse. *Neuron* **23**, 645–648 (1999).
- Verkerk, A. J. 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).
- Lines, M. A. et al. Haploinsufficiency of a spliceosomal GTPase encoded by *EFTUD2* causes mandibulofacial dysostosis with microcephaly. *Am. J. Hum. Genet.* **90**, 369–377 (2012).
- Bernier, F. P. et al. Haploinsufficiency of *SF3B4*, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome. *Am. J. Hum. Genet.* **90**, 925–933 (2012).
- 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. *Hum. Mutat.* **15**, 541–555 (2000).
- Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
- Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
- Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
- Walker, R. L. et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* **179**, 750–771.e22 (2019).
- Zhou, J. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
- Ule, J. et al. CLIP identifies Nova-Regulated RNA networks in the brain. *Science* **302**, 1212–1215 (2003).
- Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Kim, Y. K. & Maquat, L. E. UPFRONT and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. *RNA* **25**, 407–422 (2019).
- Hogg, J. R. & Goff, S. P. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell* **143**, 379–389 (2010).
- Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
- Larsen, T. M. et al. Family history of psychiatric illness as a risk factor for schizoaffective disorder: a Danish register-based cohort study. *Arch. Gen. Psychiatry* **62**, 841–848 (2005).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
- Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
- Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
- Ruderfer, D. M. et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).
- Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
- Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
- 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).
- 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. *Am. J. Hum. Genet.* **105**, 509–525 (2019).
- Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
- Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
- Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
- Lam, M. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
- Schork, A. J. et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
- Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
- Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
- 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. *Nat. Genet.* **51**, 245–257 (2019).
- Howard, D. M. et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* **9**, 1470 (2018).
- de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
- Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
- Shi, Y. et al. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat. Genet.* **43**, 1224–1227 (2011).
- Åberg, K. et al. Human QKI, a new candidate gene for schizophrenia involved in myelination. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **141B**, 84–90 (2006).
- Bhalala, O. G., Nath, A. P., Inouye, M., Sibley, C. R. & UK Brain Expression Consortium Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. *PLoS Genet.* **14**, e1007607 (2018).
- Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Inloes, J. M. et al. The hereditary spastic paraplegia-related enzyme DDHD2 is a principal brain triglyceride lipase. *Proc. Natl Acad. Sci. USA* **111**, 14924–14929 (2014).

51. Finkel, R. S. et al. Nusinersen versus sham control in infantile-onset spinal muscular atrophy. *N. Engl. J. Med.* **377**, 1723–1732 (2017).
52. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).
53. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
54. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
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 Med.* **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).
57. Darnell, R. B. & Posner, J. B. Paraneoplastic syndromes involving the nervous system. *N. Engl. J. Med.* **349**, 1543–1554 (2003).
58. Häslér, R. et al. Alterations of pre-mRNA splicing in human inflammatory bowel disease. *Eur. J. Cell Biol.* **90**, 603–611 (2011).
59. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Deep learning inference of RBP dysregulation variant effects. We utilized deep convolutional neural networks (CNNs)⁶⁰ to build a quantitative model of RNA sequence features required for RBP binding, as biochemically assayed by CLIP⁶¹ (training data). These RBP models subsequently enabled probabilistic inference of the effect of sequence variants, 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; 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 (that is, no additional training or manual parameter changes). All RBPs modeled in this study showed 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 are conserved in humans⁶; 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 and C across the 1,000 base pair (bp) window anchored around the RBP binding site. The subsequent convolution layer looks at an 8 bp window that shifts by 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. Analogs to neurons, we then applied a rectifier activation function (ReLU) that sets the convolution layer output to a scale minimum of 0 (that is, ReLU(x) = max(0, x)). Thus formally, input S results in convolution layer output location n for kernel k as the following:

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

where I is the window size and D is the input depth (for example, for the first convolution layer D represents the four RNA bases and are indexed window location by i and base by d).

Next, we added 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 (second), ReLU, pooling and convolution layer (third) where higher-order sequence motifs can be derived based on the first layer local motifs (second convolution layer, 320 kernels; third convolution layer, 480 kernels with identical ReLU and pooling layer).

Lastly, we added a fully connected layer that can now take the resulting output from the three convolution steps and integrated it 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 (for example, RNA poly(A) signal, splice sites and branch points). 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 took the absolute predicted probability differences between the two alleles (reference versus alternative) computed by the CNN 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 the [0,1] scale for fitting the 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 gnomAD¹⁹ cohort variants, all passing the random forest gnomAD quality filter, were downloaded and filtered for noncoding region variants (that is, non-repeat regions of 5' UTR, intron and 3' UTR) of protein-coding genes. (We used 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 (reference versus alternate allele) across all variants were standardized to have an s.d. of 1 to obtain the final RBP dysregulation estimates. The gnomAD cohort allele frequencies were used to segregate the variants into four different MAF 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 subgenic regions were conducted as described previously¹⁶, annotating to 5' UTR, 3' UTR or 200 bp introns flanking an exon previously observed to be alternatively spliced⁶². The RBP subgenic selection signature was assessed by fitting a linear model regressing RBP dysregulation levels on allele frequency and subgenic annotations and evaluating the statistical significance of the interaction term between the two explanatory variables. This

essentially involved 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 subgenic selection signatures with an FDR < 0.05 for only 3' UTR or 5' UTR regions.

Estimating effect sizes due to RBP dysregulation. The extensive LD between SNPs in the human population provides 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 used the previously published statistical framework of stratified LD score regression²⁴ 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 (for example, RBP dysregulation level or coding SNP), r_{jk} is the correlation between SNPs j and k in the reference panel (selected to best match the GWAS cohort) and b measures the confounding bias²⁴. Lastly, τ_c and the final standardized form τ'_c —normalized by the total SNP-based heritability and s.d. of an annotation—represent the estimated effect size of the annotation²⁵:

$$\tau'_{\text{RBP}} = \frac{M^* \text{s.d.}_{\text{RBP}}}{h^2} \tau_{\text{RBP}}$$

More formally for RBP dysregulation annotations, τ' represents the per-SNP heritability ($\frac{h^2}{M}$, M being the number of common SNPs) associated with an s.d. increase of variant RBP effect (s.d._{RBP}). We restricted our RBP predictions to SNPs from the 1000 Genomes Project (European cohort) and fitted the regression model only on HapMap SNPs with an MAF > 0.05 as conducted previously²⁴. The block jackknife procedure was used to test statistical deviation from zero for each fitted τ' .

As presented in the regression model, we fitted τ'_{RBP} by conditioning on a collection of baseline annotations to avoid upward bias in the effect size estimation. We obtained the collection of baseline annotations previously used in the stratified LD score regression study (that is, baseline LD)²⁵, which 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, which collectively results in appropriately calibrated null uniform RBP P values based on the permutation test shown in Supplementary Fig. 8. Additional baseline annotations included nonfunctional annotations, such as allele age, MAF, low levels of LD, CpG content and background selection statistics. We excluded conservation-based annotations since 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 (the full 71 baseline annotations are listed in Supplementary Table 5; the baseline annotation correlation with RBP annotations is shown in 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 1000 Genomes Project European reference panel using simGWAS⁶⁶. 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 Consortium⁶⁷, and both promoters⁶⁸ and coding regions (restricted to nonsynonymous variants) expressed in the brain⁶⁹. For each causal SNP effect size, we modeled using a Fisher polygenic model with trait heritability set to $h^2 = 0.5$. Each simulated GWAS was fitted 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 PGC, which were uniformly processed and analyzed and were sufficiently powered to observe genome-wide significant SNPs. Datasets on five

disorders met these criteria—ADHD, ASD, 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 cohort³⁸ and Danish cohort from the iPSYCH consortium cross-disorder GWAS study³⁹ were used for the replication analysis. For cross-disorder replication comparison, the non-iPSYCH-overlapping PGC cohorts obtained from older PGC GWAS studies were used for cross-disorder RBP effect size estimates^{32,70–73}. The population-matched schizophrenia versus bipolar disorder case GWAS³⁰ summary statistic was obtained from the PGC Web portal. (The full list of GWAS studies examined in this work can be found in Supplementary Table 6.)

Joint modeling of molecular QTLs. Fine-mapped GTEx²⁶ expression QTL (eQTL) (FE-meta) and BLUEPRINT⁴⁹ molecular QTL annotations were obtained from a previous study³⁵ that generated and validated the maximum causal posterior probability (MaxCPP)-based QTL annotations for GWAS enrichment analysis. The CommonMind⁷⁴ isoform QTLs (isoQTLs) and GTEx sQTL (brain cortex version 8) were fine-mapped to produce MaxCPP annotations following the same procedure as reported previously³⁵. 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 expressed in the human prefrontal cortex with developmental stage and age from Jaffe et al.⁷⁵. Peak childhood stage annotations were obtained by filtering for regions that showed both differential and maximum expression during childhood compared to other age groups.

To estimate the collective RBP dysregulation effects compared to other annotations (for example, sQTLs), we leveraged the out-of-sample East Asian schizophrenia GWAS³⁸. To find the top combined RBP dysregulation effect SNPs associated with schizophrenia, we jointly fitted all RBP models in the regression model to SNP chi-squared values from the East Asian GWAS via LD score regression. Next we used the joint RBP regression model (that is, weighting each RBP model effects by the estimated regression coefficients) to find the top combined RBP dysregulation effect SNPs (top 0.1, 0.5 or 1%) to examine the amount of heritability they capture compared to other annotations (for example, sQTL) in the PGC European schizophrenia GWAS (that is, nonoverlapping cohort with the model fitting the East Asian GWAS).

Genetic architecture analysis. LOF-intolerant genes were obtained by the ExAC consortium with a pLI threshold of >0.9 as described previously¹⁶. For the stratified LD score regression models, we jointly fitted, for each RBP model, the LOF-intolerant and non-LOF-intolerant gene region 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 included SNP to LOF-intolerant gene regions and SNP to LOF-intolerant gene coding regions to prevent potential upward bias due to the general higher background heritability enrichment levels. We also added the two SNP to LOF-intolerant genes or their coding region baseline annotations for the differential risk analysis between schizophrenia and bipolar disorder to mitigate any potential bias.

MAGMA (v.1.07b)⁴⁴ was used to estimate the gene-level association with schizophrenia and bipolar disorder. GENCODE⁷⁶ v.25 gene annotations lifted to GRCh37 coordinates and a total of 19,984 protein-coding genes were analyzed. We used SNPs from 10,000 upstream of the gene body and 1,500 downstream for each gene as used previously in a PGC schizophrenia GWAS analysis study⁷⁷ along with the 1000 Genomes Project²⁷ European reference panel. Colocalization analysis for the *ILF3* locus between schizophrenia and bipolar GWAS³⁰ and GTEx ILF3 eQTL (v.8 meta-tissue) was conducted using Coloc (v.3.2)⁷⁸. The ILF3 TWAS results were obtained from the psychENCODE study⁴⁵.

Genetic correlation analysis stratified by RBP dysregulation was conducted by first fitting a joint RBP regression model for each disease/condition to estimate the collective phenotypic effect of RBP dysregulation for each SNP. Next, when examining a pair of phenotypes (x, y), we iterated over each independent LD block⁷⁹ and extracted the top RBP dysregulation variant that influenced disorder x and their GWAS effect sizes on both x and y . We then calculated the correlation between the GWAS effect sizes on x and the GWAS effect sizes on y .

Functional mapping of *DDHD2*. QKI eCLIP and knockdown RNA sequencing data were obtained by the ENCODE project⁴⁹ for the K562 and HepG2 cell lines. eCLIP data were processed as described previously¹⁶ and visualized in the Integrative Genomics Viewer (v.2.4.13)⁸⁰. Kallisto (v.0.44.0)⁸¹ coupled with Sleuth (v.0.30.0)⁸² was used for differential expression analysis of the *DDHD2* transcript (ENST00000520272) after QKI knockdown. P values were computed using the likelihood ratio test implemented in Sleuth and the FDR was computed across all transcripts using Storey's q (ref. ⁸³). The genotyping results for SNP rs6981405 in the K562 and HepG2 cell lines were obtained from the ENCODE project.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 used in this study is available at <https://hb.flatironinstitute.org/seqweaver/about>.

References

- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
- Licatalosi, D. D. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
- Yan, Q. et al. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl Acad. Sci. USA* **112**, 3445–3450 (2015).
- Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631–1637 (2003).
- Lebedeva, S. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell* **43**, 340–352 (2011).
- Lovci, M. T. et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013).
- Fortune, M. D. & Wallace, C. simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics. *Bioinformatics* **35**, 1901–1906 (2018).
- Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
- Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
- Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
- Anney, R. J. L. 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. *Mol. Autism* **8**, 21 (2017).
- Ripke, S. et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
- Neale, B. M. et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 884–897 (2010).
- Sklar, P. et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
- Hoffman, G. E. et al. CommonMind consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Sci. Data* **6**, 180 (2019).
- Jaffe, A. E. et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
- Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
- Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
- Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2015).
- Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
- Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).
- Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).

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 National Institutes of Health grant nos. R01HG005998, U54HL117798 and R01GM071966, U.S. Department of Health and Human Services grant no.

HHSN272201000054C and Simons Foundation grant no. 395506. O.G.T. is a senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research. We thank the Simons Foundation Autism Research Initiative, Simons Foundation and Flatiron Institute. A substantial portion of the work in this paper was performed at the Terascale Infrastructure for Groundbreaking Research in Science and Engineering 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.

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., C.L.T. and R.B.D. 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

The authors declare no competing interests.

Additional information

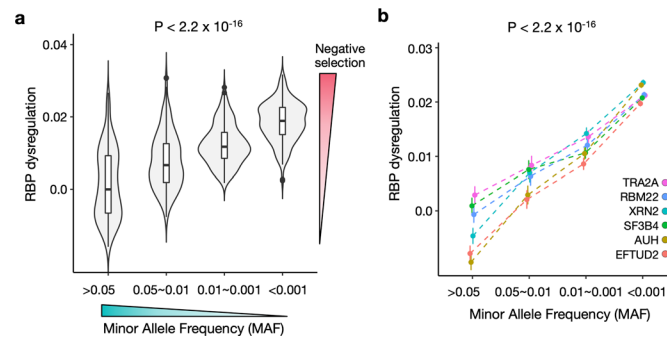
Extended data is available for this paper at <https://doi.org/10.1038/s41588-020-00761-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-020-00761-3>.

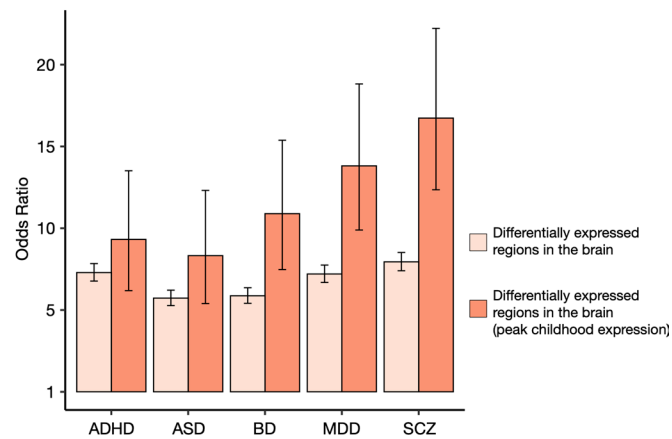
Correspondence and requests for materials should be addressed to C.Y.P. or O.G.T.

Peer review information *Nature Genetics* thanks Amalio Telenti and Thomas Werge for their contribution to the peer review of this work.

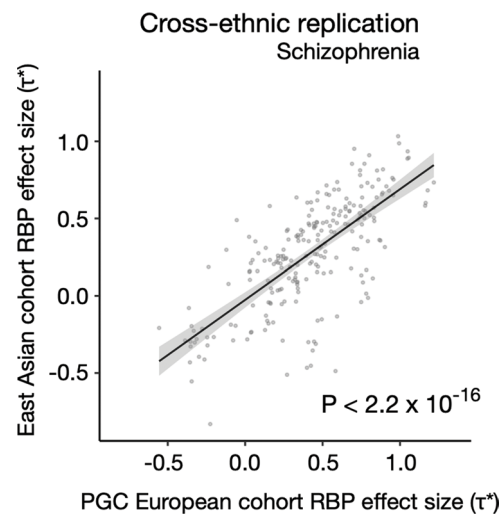
Reprints and permissions information is available at www.nature.com/reprints.



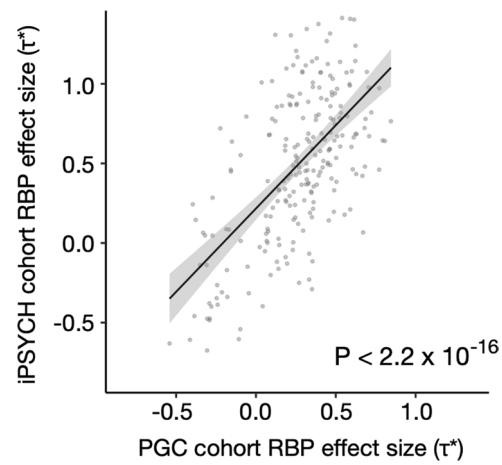
Extended Data Fig. 1 | Population genetics reveals negative selection acting on RBP target site dysregulation. a, Across the Seqweaver profiled RBPs, we observe differential selection signatures for variants when segregated by their RBP target site dysregulation levels. Specifically, for gnomAD cohort 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).



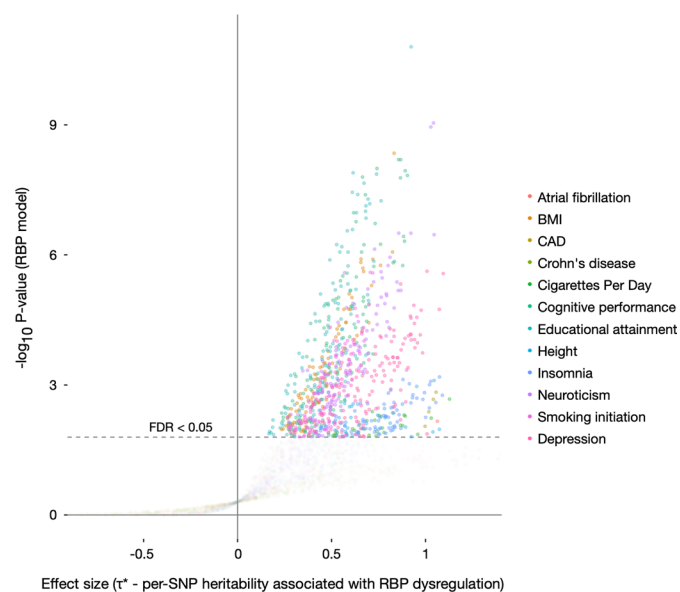
Extended Data Fig. 2 | Regions with peak childhood stage expression shows the largest enrichment association with RBP dysregulation. We test the overlap between prefrontal cortex brain differentially expressed regions and RBP dysregulation SNPs (the top 0.5%) associated with each disorder in comparison 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.



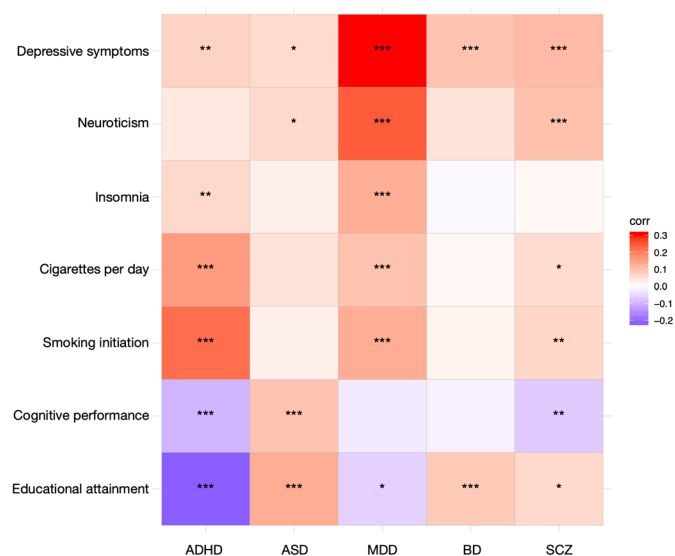
Extended Data Fig. 3 | Cross-ancestry replication – RBP dysregulation effects replicate in an independent cohort. Replication of estimated schizophrenia RBP dysregulation effect sizes (τ^* , European Psychiatric Genomics Consortium (PGC)) when compared to estimates from an East Asian cohort (Lam et al). P-value computed using spearman rank test of RBP effect sizes.



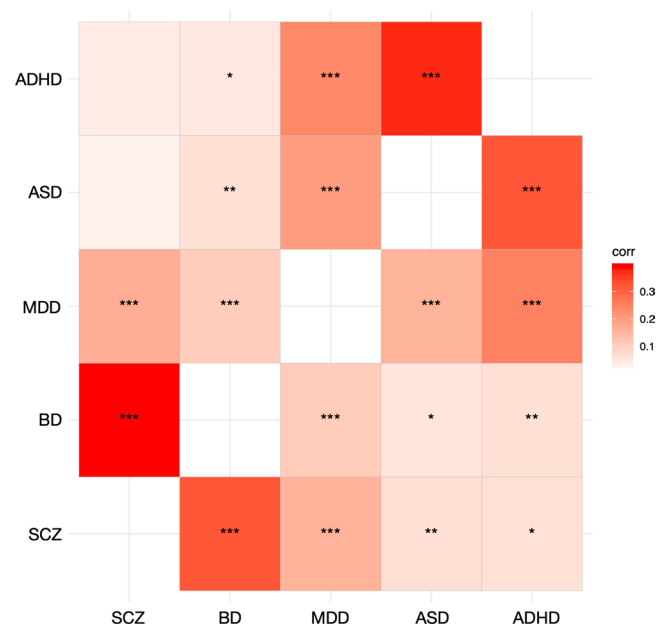
Extended Data Fig. 4 | RBP dysregulation effects for cross-disorder risk replicate in iPSYCH cohort. Replication of estimated cross-disorder RBP dysregulation effect sizes (τ^* , Psychiatric Genomics Consortium cohort) when compared to estimates from the iPSYCH cohort. P-value computed using spearman 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).

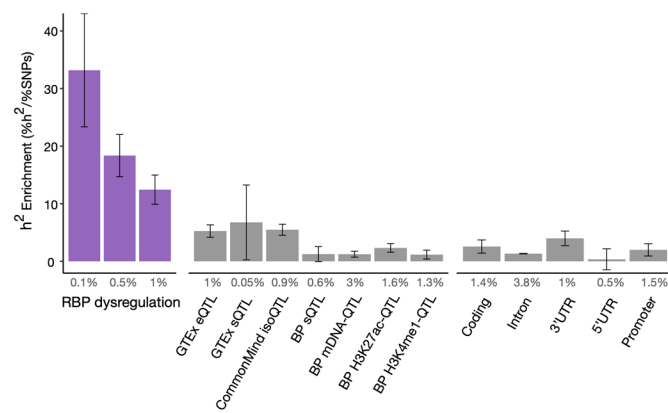


Extended Data Fig. 6 | Heatmap showing patterns of correlated GWAS effect sizes between psychiatric disorders and behavioral-cognitive phenotypes for variants affecting RBP dysregulation. For each pair of disorder and phenotype (x,y), we extracted the top RBP dysregulation set of variants that influence 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 *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

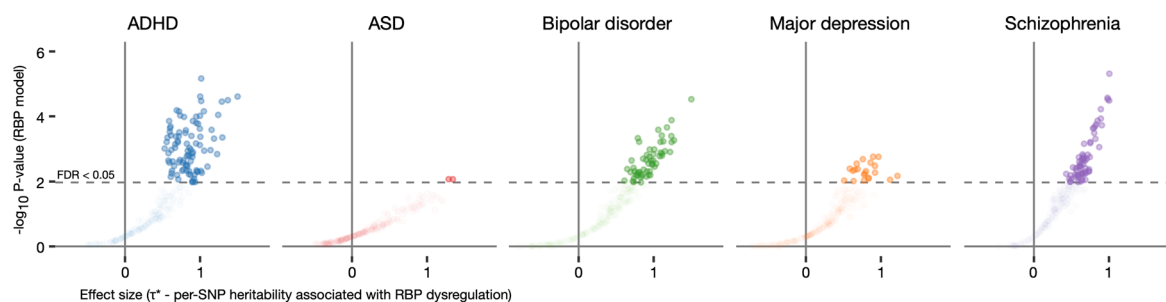


Extended Data Fig. 7 | Heatmap showing patterns of correlated GWAS effect sizes between psychiatric disorders for variants affecting RBP dysregulation.

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).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

no software was used.

Data analysis

Seqweaver - <https://hb.flatironinstitute.org/seqweaver>
R - 3.6 for statistically analysis including multiple hypothesis correction
LD Score Regression (LDSC) - v1.0
simGWAS - 0.2 for GWAS simulations

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All predicted variants effects can be downloaded at <https://hb.flatironinstitute.org/seqweaver>