nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-019-0382-7

A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data

Quan Wang[1,2,11], Rui Chen[1,2,11], Feixiong Cheng[3,4,5], Qiang Wei[1,2], Ying Ji[1,2], Hai Yang[1,2], Xue Zhong[2,6], Ran Tao[2,7], Zhexing Wen[8], James S. Sutcliffe[1,2], Chunyu Liu[9], Edwin H. Cook[10], Nancy J. Cox[2,6] and Bingshan Li[1,2]*

Genome-wide association studies (GWAS) have identified more than 100 schizophrenia (SCZ)-associated loci, but using these findings to illuminate disease biology remains a challenge. Here we present integrative risk gene selector (iRIGS), a Bayesian framework that integrates multi-omics data and gene networks to infer risk genes in GWAS loci. By applying iRIGS to SCZ GWAS data, we predicted a set of high-confidence risk genes, most of which are not the nearest genes to the GWAS index variants. High-confidence risk genes account for a significantly enriched heritability, as estimated by stratified linkage disequilibrium score regression. Moreover, high-confidence risk genes are predominantly expressed in brain tissues, especially prenatally, and are enriched for targets of approved drugs, suggesting opportunities to reposition existing drugs for SCZ. Thus, iRIGS can leverage accumulating functional genomics and GWAS data to advance our understanding of SCZ etiology and potential therapeutics.

For most complex diseases, translating GWAS findings to uncover their underlying biological mechanisms and clinical applicability remains a great challenge[1]. While drug development guided by genetic evidence should have greater rates of success[2], few effective drug targets have been identified through GWAS analysis thus far. SCZ represents a paradigmatic example of this challenge. The Psychiatric Genomics Consortium has assembled tens of thousands of samples worldwide and reported 108 genomic loci associated with SCZ in a milestone paper[3]. However, only one recognized drug target, the dopamine receptor D2 (encoded by *DRD2*), stood out from the results. The therapeutic impasse is largely a consequence of the paucity of novel targetable genes that can be accurately manipulated with drugs[4]. For most, if not all, GWAS loci it is non-trivial to pinpoint the corresponding risk genes, as the loci usually cover multiple candidate genes and the genuine risk gene (or genes) may be megabases away from the index single nucleotide polymorphisms (SNPs)[5]. Genes closest to index SNPs were intuitively assigned as risk genes in previous studies[6]; however, there is increasing evidence to suggest that risk genes may not be those in closest proximity to index SNPs[7].

There has been tremendous effort in the past few years to dissect the machinery of gene regulation. Epigenomics data generated in large-scale projects such as Functional Annotation of the Mammalian Genome 5 (FANTOM5) provide critical links between regulatory elements and the genes that they regulate[8]. Moreover, the recent advances of genome-scale chromosome conformation capture (Hi-C) technology provide global views of both short- and long-range interactions among genomic loci[9,10]. Hi-C data have been successfully used to infer the long-range interaction between distal regulatory elements (DREs) and target promoters[9,10]. These studies showed the promise of linking GWAS loci to disease risk genes; however, such data accumulated to date are far from satisfactory. Meanwhile, individual omics data provide complementary support so that integrating multi-omics data is expected to strengthen the signal for pinpointing risk genes. At a different level, multi-omics data on individual risk genes are further amplified when multiple risk genes are considered together, particularly given the polygenicity of diseases such as SCZ and that disease risk genes often converge on related biological processes. Intuitively, the increased precision achieved by using the joint modeling approach is due to borrowing supporting evidence from risk genes across all GWAS loci.

In this study, we developed a Bayesian framework, entitled iRIGS, to probabilistically infer risk genes driving GWAS signals by integrating the following two layers of information: (1) multiple lines of supporting evidence from multi-omics data for individual genes, and (2) relationships of genes in the biological networks. In its simplest form, the framework can be viewed as a Bayesian model selection problem; that is, to select genes from each of the GWAS loci such that the supporting evidence on the selected risk genes from all loci is collectively high. The proposed method is flexible in that it can leverage data from different sources (for example, transcriptomics and epigenomics data) and cumulatively orchestrate them to calculate the probability for risk gene prediction. The application of iRIGS to SCZ showed that our predicted risk genes explain significantly enriched heritability, are highly consistent with the leading pathophysiological hypotheses of SCZ, and are

[1]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA. [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [3]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [4]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA. [5]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA. [6]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [7]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA. [8]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, USA. [9]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA. [10]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA. [11]These authors contributed equally: Quan Wang and Rui Chen. *e-mail: bingshan.li@Vanderbilt.Edu

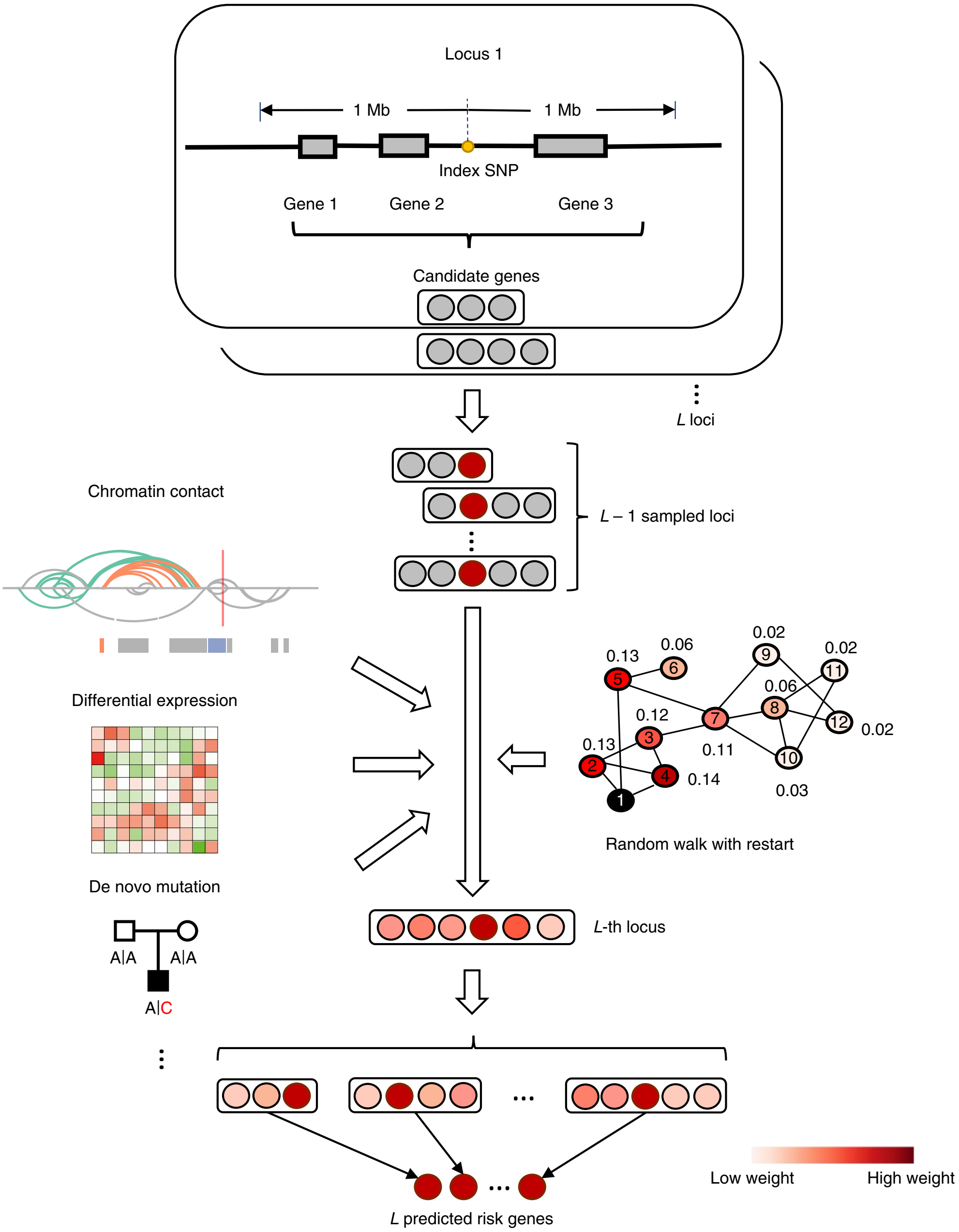

Fig. 1 | Schematic of the iRIGS framework. Each circle represents a candidate gene, and candidate genes from a GWAS locus are arranged horizontally. Candidate genes from different GWAS loci are stacked vertically. In the middle of the figure, the risk genes for the (L – 1) loci have already been sampled, and for the L-th locus, the colors of the genes represent the strength of the support from genomic features as well as the closeness to the (L – 1) sampled risk genes in the network space. After the sampling converges, the candidate gene with the highest PP at each locus is denoted as the inferred risk gene.

significantly enriched in targets of approved drugs. Taken together, these results confirm and greatly expand our previous understanding of the disease biology of SCZ and support the ability of our framework in translating GWAS findings into biological mechanisms and clinical applicability.

Results

Overview of the iRIGS framework. Figure 1 provides a schematic of the framework. Let L denote the number of GWAS loci, and for a specific locus we collected all the genes located within a 2 Mb region centered at the index SNP as its candidates. The goal of iRIGS is to probabilistically rank candidate genes at each GWAS locus based on their cumulative supporting evidence and closeness in a gene–gene network. Specifically, our goal is to find a set of L genes, each selected from one GWAS locus, such that the selected L genes achieve the highest score underlying a specified scoring scheme. Computationally, it is infeasible to enumerate all possible gene combinations; therefore, we adopted a Gibbs sampling algorithm to address the challenge, transitioning the problem into a conditional single-dimensional sampling procedure. For example, when sampling the risk gene from candidates at the L-th locus, we assume that the risk genes at all other (L – 1) loci have been selected, and the sampling probability for a gene at the L-th locus is computed conditionally on the (L – 1) risk genes based on the combined supporting evidence from this gene's multi-omics data as well as its closeness to the other (L – 1) risk genes in the network. The sampled

Table 1 | Enrichment of NRGs and HRGs in gene sets implicated in SCZ

Gene set[a]	NRG vs LBG		HRG vs WBG		HRG vs LBG	
	$P_{\text{corrected}}$	OR	$P_{\text{corrected}}$	OR	$P_{\text{corrected}}$	OR
AutDB (781)	1.23×10^{-8}	10.75 (18)	2.87×10^{-14}	9.04 (27)	4.20×10^{-16}	18.22
ECG (998)	1.60×10^{-4}	4.63 (17)	3.21×10^{-16}	8.85 (32)	9.69×10^{-15}	10.65
Essential genes (3,910)	4.19×10^{-11}	4.91 (48)	9.23×10^{-8}	3.35 (46)	3.00×10^{-9}	4.25
FMRP-Darnell (832)	1	1.85 (9)	3.28×10^{-9}	6.42 (22)	5.95×10^{-8}	6.86
RBFOX1 (556)	0.11	3.60 (8)	4.98×10^{-4}	4.71 (12)	2.36×10^{-5}	9.10
miR-137 targets (281)	4.24×10^{-5}	9.79 (11)	3.29×10^{-5}	7.82 (10)	5.69×10^{-5}	11.18
PSD (1,444)	4.03×10^{-5}	4.52 (20)	2.21×10^{-3}	2.94 (19)	8.74×10^{-5}	4.42
FMRP-Ascano (939)	0.41	2.38 (10)	1.55×10^{-3}	3.51 (15)	4.30×10^{-3}	3.61
CCS (73)	1	2.03 (1)	6.57×10^{-4}	14.94 (5)	4.38×10^{-3}	21.34
PRAZ (209)	1	2.04 (2)	1.82×10^{-3}	7.14 (7)	4.78×10^{-3}	8.69
mGluR5 (37)	1	0 (0)	0.02	17.60 (3)	0.08	25.13
PRP (336)	1	2.76 (4)	0.52	3.03 (5)	1	2.48
TADA (179)	1	4.10 (2)	1	2.22 (2)	1	3.32
ARC (25)	1	Inf (1)	1	8.16 (1)	1	Inf
PSD-95 (107)	1	8.20 (2)	1	3.75 (2)	1	8.31
NMDAR (59)	0.60	16.39 (2)	1	3.37 (1)	1	8.24
SYV (107)	1	2.73 (2)	1	1.84 (1)	1	2.06
GABA_A (18)	1	0 (0)	1	0 (0)	1	0

[a] The numbers of genes in the corresponding gene sets are in parentheses. One-sided Fisher's exact test and Bonferroni correction were used for enrichment analyses. Please refer to the Methods for details of gene set abbreviations.

gene at the L-th locus is then put back into the set of selected genes, and the sampling for the (L – 1)-th locus is iterated until all the loci are visited. This process is repeated until risk genes converge on a stationary distribution. The posterior probability (PP) of each candidate being a risk gene can be assessed on the basis of the sampling frequency. For each GWAS locus, one or potentially more risk genes can be selected according to the PP. In this study, we only selected one risk gene with the highest PP for each locus. Details of iRIGS can be found in the Methods and the Supplementary Note.

Applying iRIGS to identify disease risk genes for SCZ. In a milestone paper of a GWAS of SCZ, the Psychiatric Genomics Consortium reported 108 independent, genome wide-significant loci[3]. We performed iRIGS on these loci to identify risk genes for SCZ (Supplementary Table 1). A key component of iRIGS is a collection of genomic features that can be used to characterize SCZ risk genes. Given our limited knowledge of the disease genes of SCZ, genuine characterization of genomic features of SCZ risk genes has not been clearly established. Here, we adapted iRIGS in a special form to infer preliminary SCZ risk genes to carry out an ab initio discovery of associated genomic features. Specifically, we used a generic gene–gene network constructed from Gene Ontology (GO) to run iRIGS without genomics data (Methods), and denoted the identified risk genes as network-derived risk genes (NRGs). In total, we predicted 104 NRGs after merging the overlapping genes across loci.

To show that these 104 NRGs harbor genuine SCZ risk genes, we assessed the enrichment of NRGs with 18 gene sets that have been widely and repeatedly implicated in SCZ (Methods). For the enrichment analysis, we selected the 842 genes with PP values less than the median PP of all candidate genes as the background and termed them local background genes (LBGs). We observed significant enrichments in five gene sets after Bonferroni correction (Table 1), including postsynaptic density (PSD) proteins ($P_{\text{corrected}} = 4.03\times10^{-5}$, odds ratio (OR) = 4.52) and microRNA-137 (miR-137) targets ($P_{\text{corrected}} = 4.24\times10^{-5}$, OR = 9.79). We also tried other thresholds to define LBGs, and found that the enrichment patterns were similar (Supplementary Table 2). Thus, in subsequent analyses, we used the median PP as the LBG threshold. To further determine that the enrichment is not biased owing to GO annotations, we applied iRIGS to two other traits (age-related macular degeneration and obesity) using the same GO network. No enrichments were observed in any of these gene sets ($P_{\text{corrected}} > 0.05$).

We also evaluated whether NRGs carry more de novo mutations (DNMs) identified in SCZ compared to LBGs. We collected DNM data of parent–proband trios as well as unaffected siblings from previous studies[11] (Methods). We only focused on the predicted deleterious DNMs (pdDNMs), defined as loss-of-function (nonsense, splicing, and frame shift) DNMs or missense DNMs with a deleterious score (DScore) of >3, in which DScore is defined as the number of deleteriousness predictions among 12 algorithms reported by ANNOVAR[12] (Methods). We observed a significant enrichment of proband pdDNMs with NRGs ($P = 2.53\times10^{-3}$, OR = 3.88), while no enrichment was observed for synonymous DNMs ($P = 1$). By contrast, no significant enrichments were observed for either pdDNMs ($P = 1$) or synonymous DNMs ($P = 1$) identified in unaffected siblings.

Discovery of characteristic genomic features of SCZ risk genes. These preliminary explorations showed that the predicted NRGs capture the genetic risks of SCZ. We therefore used the 104 NRGs to explore genomics data and to learn genomic features that are characteristics of SCZ risk genes. First, we found that NRGs are more likely to exhibit differential expression (DE) compared to LBGs ($P = 8.55\times10^{-3}$) (Fig. 2a) in the CommondMind data[13] (Methods). We next explored DRE–promoter links of NRGs by testing the hypothesis that risk genes have more incoming regulatory links compared to background levels. We found that NRGs are indeed connected to more DREs in the Hi-C and FANTOM5 data (Fig. 2b, Supplementary Fig. 1, Methods, and Supplementary Note). We also

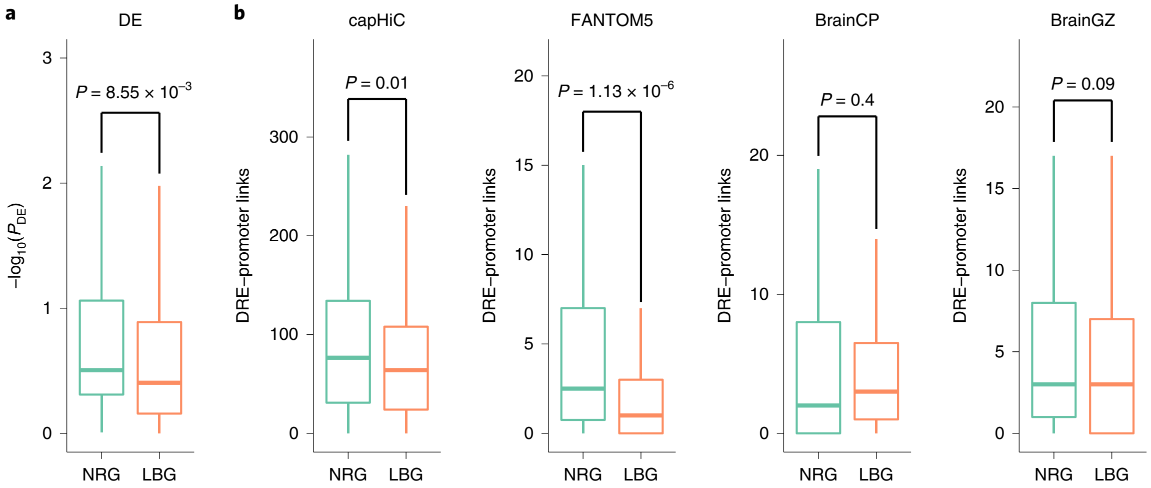

Fig. 2 | Discovery of genomic features characteristic of SCZ risk genes. **a**, NRGs are more likely to exhibit DE compared to LBGs. We directly used the *P* values of DE from the CommondMind Consortium to perform the comparison (one-sided Wilcoxon rank-sum test, n = 99 and 562 for NRGs and LBGs, respectively). **b**, NRGs capture more DRE-promoter links based on the data from capture Hi-C (capHiC), FANTOM5, and brain-specific Hi-C (cortical and subcortical plate (BrainCP) and germinal zone (BrainGZ)). One-sided Wilcoxon rank-sum test. For capture Hi-C and FANTOM5, n = 104 and 842 for NRGs and LBGs respectively; for brain-specific Hi-C, n = 104 and 831 for NRGs and LBGs respectively. See the main text and Supplementary Note for details. The box plots show the median and the 25th and 75th percentiles. The whiskers extend from the box to the largest and smallest values no further than 1.5 times the inter quartile range (IQR) from the box (or the distance between the 25th and 75th percentiles). NRGs, network-derived risk genes; LBGs, local background genes.

investigated the distance between NRGs and index SNPs. Although variants identified in GWAS do not necessarily implicate the nearest genes[7], this will nevertheless be the case for a number of risk loci. Among the 104 NRGs we predicted, 23 genes (22%) were the nearest ones to the corresponding index SNPs, significantly higher than expected (P = 0.04, permutation test).

Integrating the learned genomic features to identify high-confidence risk genes for SCZ. As shown above, different genomic features (that is, DNMs, DE, DRE–promoter links, and distance to index SNP (DTS)) consistently exhibited supportive evidence for NRGs. We therefore integrated them into iRIGS (Methods) and predicted a total of 104 high-confidence risk genes (HRGs) (Supplementary Table 1). We next evaluated whether and how the integrated multidimensional genomic features can improve our prediction.

Genomic features show aggregated effects on HRGs. For each GWAS locus, we calculated the ratio of maximum and median PP values of local candidate genes and found that HRGs carry significantly higher sampling probabilities than NRGs ($P = 1.02 \times 10^{-18}$) (Fig. 3a). This result demonstrates the strong and aggregated influence of multidimensional genomic features on nominating risk genes.

Genomic positions of HRGs relative to GWAS index SNPs. Among the 104 HRGs, 39 genes (38%) were nearest to the corresponding GWAS index SNPs (16 more genes than NRGs), significantly higher than expected ($P < 1 \times 10^{-6}$, permutation test). The extreme significance strongly supports the effectiveness of incorporating genomic features in selecting genuine risk genes. In particular, HRGs that are also the nearest genes to index SNPs provide high-confidence candidates for follow-up studies.

Of particular interest are the remaining 65 HRGs that are not the nearest genes to the index SNPs. For each of these 65 HRGs, denoted as non-nearest HRGs, we picked the nearest gene from the corresponding locus as a control. We also performed a gene set enrichment analysis to compare the 65 nearest non-HRG genes with the 65 non-nearest HRGs. The gene sets used here for this analysis are the same as the ones presented in Table 1. We found that the non-nearest HRGs are more significantly enriched in the gene sets compared to the nearest non-HRG genes (Supplementary Table 3), suggesting that the 65 non-nearest HRGs identified by iRIGS are more likely to be true risk genes than their nearest counterparts.

In the 2 Mbp window of GWAS index SNPs, most candidate genes were out of the linkage disequilibrium (LD) blocks of the index SNPs. We therefore investigated the extent to which the identified HRGs are in GWAS loci LD blocks, which were defined as regions with $r^2 > 0.2$ with the index SNPs. Among the 104 identified HRGs, 34 (33%) genes were in LD blocks. For 39 HRGs that were also nearest to the index SNPs, around half (19) were in LD blocks, while for the 65 non-nearest HRGs, only 15 were in LD blocks.

HRGs explain high disease heritability. We then utilized stratified LD score regression (LDSC) to evaluate the SCZ heritability explained by HRGs[14]. We included the SNPs located within a 20 kb window centered at the transcription start site of each gene for LDSC analysis. We observed that HRGs explain significantly enriched disease heritability (enrichment = 39.36, $P = 5.56 \times 10^{-7}$) compared to LBGs (enrichment = 10.06, $P = 6.31 \times 10^{-14}$) (Fig. 3b). When only focusing on the 65 non-nearest HRGs, we also observed a significant enrichment in heritability (enrichment = 19.72, $P = 2.53 \times 10^{-4}$); however, the majority of these were not in strong LD with the index SNPs (Fig. 3b). As expected, the enrichment of nearest HRGs was the highest, since they are close to index SNPs. We also tried different window sizes around the transcription start site (from 20 kb to 200 kb) for LDSC and observed the same trend of enrichments. Note that DTS is a confounding effect for LDSC, since genes close to index SNPs are more likely to have a high LDSC score, and the HRGs used here for the LDSC analysis were obtained without the use of DTS in iRIGS.

The above evaluations demonstrate the effectiveness of incorporating multidimensional genomic features and strongly suggest that iRIGS is capable of nominating SCZ disease risk genes. In the

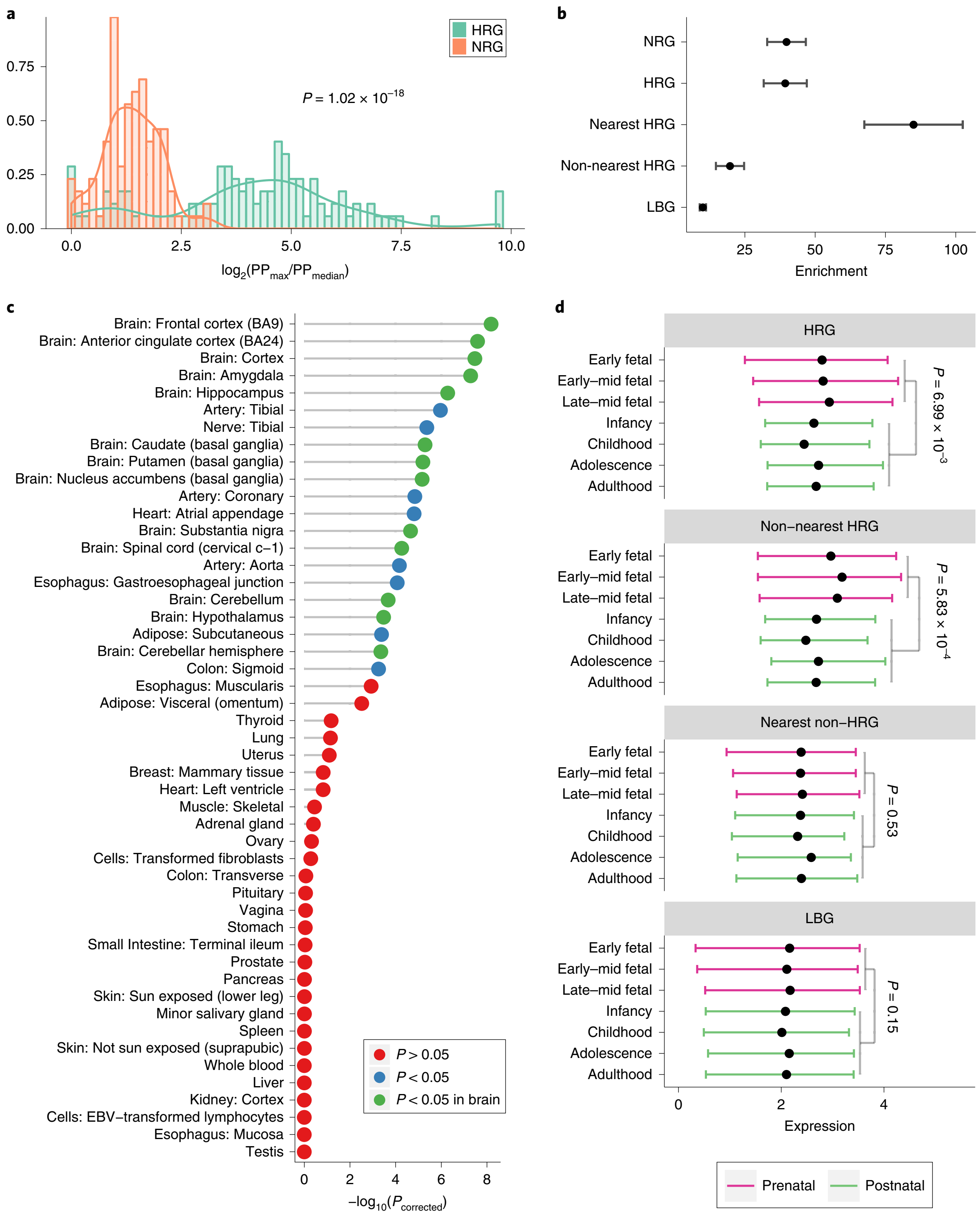

Fig. 3 | Characteristics of predicted risk genes. **a**, Distributions of the PP values of HRGs and NRGs show that HRGs carry significantly higher sampling PP values than NRGs (one-sided Wilcoxon rank-sum test, $n = 104$ for both HRGs and NRGs). The *x* axis represents the ratio of maximum and median of PP values of candidate genes for each GWAS loci. **b**, Stratified LDSC to evaluate the enrichment of SCZ heritability explained by different groups of genes. The center values represent the enrichment, and the error bars indicate standard errors. **c**, The tissue specificity of HRGs across tissues in GTEx showed that HRGs are highly expressed in brain-related tissues (one-sided Wilcoxon rank-sum test and Bonferroni correction, $n = 104$ and 830 for HRGs and LBGs, respectively). **d**, The expression of HRGs, the 65 non-nearest HRGs, the corresponding 65 nearest non-HRG genes, and LBGs across developmental stages based on the BrainSpan data show that HRGs and non-nearest HRGs are highly expressed at prenatal stages compared to postnatal stages, while the 65 corresponding nearest non-HRG genes and LBGs are not differentially expressed across developmental stages (one-sided Wilcoxon rank-sum test using medians of expression at prenatal ($n = 3$) and postnatal ($n = 4$) stages). It also shows that HRGs have higher expression levels in the brain than LBGs, consistent with the observation in **c** based on GTEx data. The error bar plot shows the median and the 25th and 75th percentiles. NRGs, network-derived risk genes; LBGs, local background genes.

Table 2 | Selected HRGs involved in biological functions implicated in SCZ

Gene	Descriptions	Nearest[c]	Refs.
Calcium channel and signaling			
CACNA1C[a]	Encodes an alpha-1 subunit of a voltage-dependent calcium channel and a target of miR-137	Yes	3
CACNB2[a]	A member of the voltage-gated calcium channel superfamily	No	3
PTK2B[b]	Involved in calcium-induced regulation of ion channels; interacts with *DAO*, a potential SCZ gene implicated from a non-GWAS signal	No	18, 19
Neurogenesis			
SOX2[a]	A transcription factor essential for neurogenesis	Yes	9
SATB2[a]	Essential for cognitive development and involved in long-term plasticity processes	Yes	20, 21
Glutamatergic neurotransmission and synaptic plasticity			
GRIA1[a]	An ionotropic glutamate receptor that mediates fast synaptic transmission	No	22
GRIN2A[a]	A glutamate-gated ion channel protein and a key mediator of synaptic plasticity; a target of miR-137	Yes	23, 24
GRM3[a]	Encodes glutamate metabotropic receptor 3, one of the major excitatory neurotransmitter receptors in the CNS; has been extensively explored as a potential drug target in SCZ	Yes	25
GPM6A[b]	Involved in neuronal plasticity and probably synapse formation; has previously been shown to be associated with the severity of depression in patients with SCZ	Yes	26
NLGN4X[b]	Involved in the formation and remodeling of CNS synapses; knockdown directly affects the neurodevelopment process, indicating a role in the molecular pathophysiology of psychiatric diseases, including ASD and SCZ	Yes	27, 28
Targets of miR-137			
TCF4[a]	Encodes transcription factor 4 and is involved in the initiation of neuronal differentiation	Yes	29–31
ZNF804A[a]	A zinc finger-binding protein previously implicated in SCZ	Yes	32
RORA[b]	Encodes a ligand-dependent transcription factor regulator; a potential ASD gene	Yes	33, 34
CSMD1[b]	A target of miR-137 and a potential SCZ gene	Yes	35

[a] Well-established SCZ genes. [b] Potential SCZ genes of great interest predicted by iRIGS. [c] Nearest (or not) to the index SNPs.

next sections, we used the 104 predicted HRGs to comprehensively characterize various properties of identified (putative) SCZ risk genes to gain further biological insights into SCZ.

Tissue-specific and developmental stage-specific expression of HRGs. We collected expression data from the Genotype-Tissue Expression (GTEx) project (Methods) and observed that HRGs have more pronounced tissue specificity in the brain than LBGs (Fig. 3c). We also observed a higher expression level of HRGs in prenatal than postnatal stages ($P=6.99\times10^{-3}$) (Fig. 3d) in the BrainSpan data[15] (Methods), while this pattern was absent for LBGs ($P=0.15$) (Fig. 3d). In addition to LBGs, we generated a set of whole-genome background genes (WBGs) by including all the human genes minus the HRGs for comparison. As expected, no difference was observed for WBGs between prenatal and postnatal stages ($P=0.27$) (Supplementary Fig. 2).

We also compared the spatiotemporal expression pattern between the 65 non-nearest HRGs and the corresponding 65 nearest non-HRG genes. We found that the non-nearest HRGs were highly expressed in prenatal stages ($P=5.83\times10^{-4}$) (Fig. 3d), while there was no significant difference for the nearest non-HRG genes ($P=0.53$) (Fig. 3d). Interestingly, for the GTEx data, we observed that the nearest non-HRG genes were highly expressed in a majority of brain tissues, while the non-nearest HRGs were not (Supplementary Fig. 3).

Involvement of HRGs in biological functions implicated in SCZ. We repeated the enrichment analysis of HRGs with the same 18 gene sets as previously used. We observed a dramatic improvement in the enrichments of SCZ-relevant gene sets compared to NRGs. Under the criterion $P_{\text{corrected}}<0.05$, NRGs were enriched in five gene sets while HRGs were enriched in ten sets, the majority of which showed remarkably enhanced ORs (Table 1). Among the ten significantly enriched gene sets, fragile X mental retardation protein (FMRP) targets, PSD, and genes related to the presynaptic active zone (PRAZ) have been extensively implicated in SCZ due to their in-depth involvement in synaptic networks. Calcium channel and signaling (CCS) is involved in multiple functions, including synaptic plasticity modulation, and has pleiotropic effects on psychiatric diseases[5]. Targets of miR-137 have been discussed in detail for the potential etiologic mechanism of SCZ[16]. Note that some of the enriched gene sets have also been previously implicated in DNM or rare coding mutation analyses, such as FMRP targets[11] and PSD[17], confirming the convergence between non-coding variants and coding mutations at the gene set level.

Table 2 lists some of the well-established SCZ genes involved in the SCZ primary functional categories derived from the aforementioned gene sets[3,9,18–35]. Specifically, *CACNA1C* and *CACNB2*, both encoding voltage-gated calcium channel subunits, are involved in CCS and contribute to the risk for SCZ[3]. *CACNA1C* is differentially expressed in patients with SCZ ($P=0.03$), and both *CACNA1C* and *CACNB2* capture multiple Hi-C links in the brain Hi-C data (Supplementary Fig. 4), contributing to the high PP of both genes in iRIGS. We also predicted two DNA-binding proteins, SOX2[9] and SATB2[20,21] (Supplementary Fig. 4), which have important roles in neurogenesis and have been widely implicated in SCZ. Several of our predicted HRGs are miR-137 target genes, including the aforementioned *CACNA1C*, and three other genes, *GRIN2A*, *TCF4*, and *ZNF804A*. *GRIN2A*, a glutamate-gated ion channel protein and a key mediator of synaptic plasticity[23,24], has a pdDNM and multiple regulatory connections (Supplementary Fig. 4). *TCF4*, which encodes transcription factor 4, participates in the initiation of neuronal differentiation by regulating the intrinsic excitability of prefrontal cortical neurons[29,30], and knockdown of *TCF4* alters the expression of genes important for developing prefrontal neocortex[29,31]. *TCF4* is also linked with numerous DREs (Supplementary Fig. 4) in iRIGS. The reduced expression of *ZNF804A* in human neurons, especially in the fetal brain, has been widely observed and hypothesized to contribute to SCZ etiology by affecting neurite growth and loss of dendritic spine density[32].

We emphasize that in addition to these well-established SCZ genes, iRIGS also nominated novel or non-canonical genes, especially genes that are distal to the index SNPs. A particular example is the rs2514218 locus, in which *DRD2*, the target of all effective antipsychotic drugs, is the nearest to the index SNP. At that locus, the top predicted gene is neural cell adhesion molecule 1 (*NCAM1*), which is distal to the index SNP, while the nearest *DRD2* is ranked third among all 16 candidate genes. We took a closer look at this region to gain more insights. *NCAM1* captured 55 Hi-C links in the brain Hi-C data (Supplementary Fig. 5), while there was only one for *DRD2*. In addition, *NCAM1* had 207 capture Hi-C links, many more than *DRD2* (111 links). We also observed four links for *NCAM1* but none for *DRD2* in the FANTOM5 data. We further explored the expression patterns of *NCAM1* and *DRD2*. In the GTEx data, *DRD2* was highly expressed in basal ganglia caudate, hypothalamus, basal ganglia nucleus accumbens, basal ganglia putamen, and substantia nigra, but the expression in cortex and frontal cortex was low (Supplementary Fig. 6). *NCAM1* was uniformly and highly expressed in all brain tissues (Supplementary Fig. 6). In the BrainSpan data, *NCAM1* showed constitutively high expression across all stages, with particularly higher expression at prenatal stages with a trajectory that peaked at the early-mid fetal stage, while *DRD2* showed lower expression across all developmental stages and no obvious pattern of transition between prenatal and postnatal stages (Supplementary Fig. 7). The spatiotemporal expression pattern of *NCAM1* is consistent with the current understanding of SCZ[36], and all these lines of evidence highlight that *NCAM1* is a promising SCZ risk gene in addition to *DRD2*. Note that the GTEx and BrainSpan data were not used in iRIGS; therefore, the spatiotemporal analysis of gene expression provides independent and unbiased support.

Another example is *PTK2B*, which is distal to the index SNP rs73229090 (the nearest gene is *CLU*). *PTK2B* encodes a kinase involved in the calcium-induced regulation of ion channels and plays an important role in regulating neuronal activity. More interestingly, *PTK2B* has been consistently found to interact with *DAO*, a potential SCZ gene implicated from non-GWAS signals[18,19]. One pdDNM and a high number of regulatory links were observed for *PTK2B* (Supplementary Fig. 8), promoting it as the top gene predicted by iRIGS. Collectively, these findings strongly indicate that *PTK2B* is a potential risk gene for SCZ.

We manually checked the remaining genes extensively, and for most of the HRGs, we found support to varying degrees for these genes to be involved in SCZ pathophysiology. We have highlighted these genes with extended supporting evidence and description in Supplementary Table 1.

Enrichment of HRGs in gene sets leading to altered neuronal phenotypes in mouse models. As it is increasingly becoming clear that SCZ reflects perturbations of neurodevelopmental processes[9,36], we were interested in assessing direct phenotypic manifestations of gene knockouts in mouse models to see whether mutations in mouse genes orthologous to HRGs exhibit phenotypes highly related to the CNS[37]. Specifically, we collected 278 gene sets relevant to CNS and behavior and neurological phenotypes from the Mouse Genome Informatics (MGI) Mammalian Phenotype Ontology (MPO) database (Methods) and observed significant enrichment of HRGs in 33 gene sets after Bonferroni correction (Supplementary Table 4). The enriched sets span from low-level molecular functions to broad behavioral phenotypes and brain morphologies, including "abnormal nervous system physiology" ($P=3.96\times10^{-7}$, OR = 6.25), "abnormal nervous system morphology" ($P=1.04\times10^{-6}$, OR = 5.23), "abnormal brain morphology" ($P=9.19\times10^{-7}$, OR = 6.86), and "abnormal behavior" ($P=3.80\times10^{-6}$, OR = 4.34).

In addition, we observed that the 65 non-nearest HRGs were significantly enriched in 19 MPO gene sets (Supplementary Table 5), while no significantly enriched gene sets were observed for the 65 nearest non-HRG genes (Supplementary Table 6). This result provides strong and orthogonal support for risk genes identified via iRIGS that are beyond the proximity to the GWAS index SNPs.

HRGs are likely to be potential drug targets. We were interested in whether the predicted HRGs have the potential for repositioning existing drugs for SCZ treatment. We curated a list of 2,263 confirmed druggable targets from multiple sources (Methods), and found that 28 (27%) HRGs are targets of 198 drugs that are approved by the US FDA, clinically investigational, or preclinical drugs (Supplementary Fig. 9; Supplementary Table 7). The overlap is a significant enrichment compared to LBGs ($P=3.83\times10^{-7}$, OR = 3.93). We observed that the 65 non-nearest HRGs were also significantly enriched in drug targets ($P=6.30\times10^{-5}$, OR = 3.78), while the degree of enrichment of the 65 nearest non-HRG genes dramatically decreased ($P=0.03$, OR = 2.13). In particular, we found that five HRGs (*GRIA1*, *GRM3*, *KCNQ5*, *CACNA1C*, and *GRIN2A*) are targets of nervous system drugs (Supplementary Fig. 9), corresponding to a significant enrichment ($P=0.01$, OR = 3.78).

One HRG, *GRM3*, which encodes the protein mGlu3, is of particular interest as it belongs to the G protein-coupled receptor family; these receptors are the targets of the majority of clinically used drugs[38]. While there has been support in the literature for a linkage between *GRM3* and SCZ[25,39,40], there have also been contrasting reports[41,42]. The results presented here provide additional evidence to support the hypothesis that *GRM3* is a SCZ risk gene. Additionally, polymorphisms in *GRM3* have been shown to correlate with cognitive performance in healthy individuals[43,44], and cognitive impairments are an area of unmet medical need in SCZ. This suggests that our results may place mGlu3 in a particularly attractive position for the development of therapeutics for SCZ.

In addition to genes such as *GRM3*, which has previously been indicated to be genetically associated with SCZ, we propose that novel genes in our list may represent new candidates for the involvement in or potential treatment of SCZ. For example, TMEFF1 and TMEFF2 are family members that exhibit DE in the brain. They are composed of transmembrane proteins that include an epidermal growth factor-like domain along with two follistatin-like domains[45–47]. The extracellular domains of these proteins can be cleaved and released from the cell surface, potentially functioning as neurotrophic factors. It has been suggested that TMEFF2 may be trophic for dopamine neurons and that it can increase dendrite length in these cells[47,48]. Cleavage of the extracellular domain of TMEFF2 is stimulated by cytokines that induce inflammation, such as interleukin-1β and tumor necrosis factor-α[49]. The fact that this protein is expressed on the cell surface suggests that it may be a candidate for drug targeting, potentially providing a completely new therapeutic strategy for SCZ and other neurological disorders.

Discussion

SCZ is a severe psychiatric disorder that is notoriously difficult to treat, which is particularly due to the poor understanding of disease etiology. Identifying risk genes at the associated loci, in our vision, is a crucial step to bridge GWAS findings and the biology of SCZ to facilitate the development of novel therapeutics. A direct benefit of this approach is drug repositioning, as risk genes shared across different diseases provide a natural lever to repurpose drugs approved for other diseases for SCZ treatment. To bridge this gap, we developed an integrative framework, iRIGS, to pinpoint risk genes from a massive pool of candidates around SCZ-associated loci by jointly modeling high-dimensional genomic features across all GWAS loci for enhanced accuracy. As a result, we provided a gene-centric view of the genetic etiology of SCZ with strong support from multiple lines of evidence. Moreover, as a proof of concept, the predicted risk genes are strongly enriched in existing drug targets, demonstrating the promise of the identified risk genes for drug repositioning for SCZ.

Our framework has a few key strengths that are worth further in-depth discussion. iRIGS jointly integrates genomic features of a set of risk genes rather than individual genes such that the weak evidence for individual risk genes is amplified by joining forces with other ones, boosting inference accuracy. A challenge for joint modeling across GWAS loci is the correlation among risk genes. In iRIGS, instead of explicitly specifying correlations among all genes, which is impractical, the correlation is derived from gene networks. The derived correlation can be viewed as a prior in a Bayesian framework for the gene–gene covariance matrix (Methods). For implementing the algorithm, we designed a Gibbs sampling strategy to achieve the following two goals: making an astronomically challenging computational problem feasible and providing a probabilistic assessment of the selected risk genes, both of which are critical. By adopting a Gibbs sampling algorithm, we transformed the high-dimensional joint modeling process into a much simpler one-dimensional problem, not only solving the computational challenge but also providing a set of risk genes with probabilistic interpretations. It is of note that although the algorithm samples one gene from each of the loci at each iteration, when zero or more than one risk gene exists at a locus, the framework is still able to rank the genes by PP values without awareness of the exact number of risk genes at each locus. For loci that do not harbor risk genes for whatever reasons, this does not pose a challenge to the robustness of the algorithm, and the sampling is distributed evenly among the candidate genes such that none of them have pronounced PP values. Considering that it is almost impossible to specify a priori the number of risk genes at each locus (being either zero or larger than one), which is also very likely to vary widely across loci, the current implementation is robust even in the presence of these challenges.

Our framework is designed to take advantages of high-dimensional genomics data, and the more relevant genomic features are included the more accurate the prediction is. The PsychENCODE project[50], for example, is actively generating various epigenomics data for psychiatric disorders, and the accuracy of risk gene predictions for SCZ will be markedly enhanced when these data are incorporated into our framework. In addition, since the genomic features of genes at individual loci are jointly modeled across all loci, the accuracy of the prediction will also be markedly improved as more loci are identified, for example, by meta-analyses of international consortia such as the Psychiatric Genomics Consortium. Moreover, the investigated genomic loci can be expanded by including sub-GWAS variants, the *P* values of which are less than a relatively loose threshold compared with the GWAS threshold. It is our expectation that with the expansion of both genomics data and discovered GWAS loci, the identification of risk genes will be greatly improved, advancing our understanding of the biology of SCZ for the ultimate goal of guiding the development of effective therapeutics. Note that the framework is equally applicable to other complex diseases, and especially suitable for diseases with large volumes of omics data, such as transcriptomics, functional genomics, epigenomics, and others. For example, data from single cell sequencing from various immune cell types can be used for autoimmune diseases. It is our hope that this framework is able to catalyze the translation of GWAS to biology and therapeutics for a variety of complex diseases.

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/s41593-019-0382-7.

Received: 16 August 2018; Accepted: 13 March 2019;
Published online: 15 April 2019

References

1. Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
4. Breen, G. et al. Translating genome-wide association findings into new therapeutics for psychiatry. *Nat. Neurosci.* **19**, 1392–1396 (2016).
5. Harrison, P. J. Recent genetic findings in schizophrenia and their therapeutic relevance. *J. Psychopharmacol.* **29**, 85–96 (2015).
6. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278–1283 (2007).
7. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
8. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
9. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
10. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
11. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
12. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
13. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
14. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
15. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
16. Schizophrenia Psychiatric Genome-Wide Association Study Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
17. Kirov, G. et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153 (2012).
18. Verrall, L., Burnet, P. W., Betts, J. F. & Harrison, P. J. The neurobiology of D-amino acid oxidase and its involvement in schizophrenia. *Mol. Psychiatry* **15**, 122–137 (2010).
19. Yang, H. C. et al. The *DAO* gene is associated with schizophrenia and interacts with other genes in the Taiwan Han Chinese population. *PLoS One* **8**, e60099 (2013).
20. Jaitner, C. et al. Satb2 determines miRNA expression and long-term memory in the adult central nervous system. *eLife* **5**, e17361 (2016).
21. Whitton, L. et al. Cognitive analysis of schizophrenia risk genes that function as epigenetic regulators of gene expression. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **171**, 1170–1179 (2016).
22. Barkus, C. et al. What causes aberrant salience in schizophrenia? A role for impaired short-term habituation and the GRIA1 (GluA1) AMPA receptor subunit. *Mol. Psychiatry* **19**, 1060–1070 (2014).
23. Thomas, K. T. et al. Inhibition of the schizophrenia-associated microRNA miR-137 disrupts Nrg1alpha neurodevelopmental signal transduction. *Cell Rep.* **20**, 1–12 (2017).
24. Weickert, C. S. et al. Molecular evidence of N-methyl-D-aspartate receptor hypofunction in schizophrenia. *Mol. Psychiatry* **18**, 1185–1192 (2013).
25. Egan, M. F. et al. Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. *Proc. Natl Acad. Sci. USA* **101**, 12604–12609 (2004).
26. Boks, M. P. et al. Do mood symptoms subdivide the schizophrenia phenotype? Association of the *GMP6A* gene with a depression subgroup. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 707–711 (2008).
27. Yan, J. et al. Analysis of the neuroligin 3 and 4 genes in autism and other neuropsychiatric patients. *Mol. Psychiatry* **10**, 329–332 (2005).
28. Shi, L. et al. The functional genetic link of *NLGN4X* knockdown and neurodevelopment in neural stem cells. *Hum. Mol. Genet.* **22**, 3749–3760 (2013).

29. Rannals, M. D. et al. Psychiatric risk gene transcription factor 4 regulates intrinsic excitability of prefrontal neurons via repression of SCN10a and KCNQ1. *Neuron* **90**, 43–55 (2016).
30. Quednow, B. B., Brzozka, M. M. & Rossner, M. J. Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. *Cell. Mol. Life Sci.* **71**, 2815–2835 (2014).
31. Hill, M. J. et al. Knockdown of the schizophrenia susceptibility gene *TCF4* alters gene expression and proliferation of progenitor cells from the developing human neocortex. *J. Psychiatry Neurosci.* **42**, 181–188 (2017).
32. Chang, H., Xiao, X. & Li, M. The schizophrenia risk gene *ZNF804A*: clinical associations, biological mechanisms and neuronal functions. *Mol. Psychiatry* **22**, 944–953 (2017).
33. Devanna, P. & Vernes, S. C. A direct molecular link between the autism candidate gene RORa and the schizophrenia candidate MIR137. *Sci. Rep.* **4**, 3994 (2014).
34. Hu, V. W., Sarachana, T., Sherrard, R. M. & Kocher, K. M. Investigation of sex differences in the expression of RORA and its transcriptional targets in the brain as a potential contributor to the sex bias in autism. *Mol. Autism* **6**, 7 (2015).
35. Kwon, E., Wang, W. & Tsai, L. H. Validation of schizophrenia-associated genes *CSMD1*, *C10orf26*, *CACNA1C* and *TCF4* as miR-137 targets. *Mol. Psychiatry* **18**, 11–12 (2013).
36. Gulsuner, S. et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
37. Pocklington, A. J. et al. Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron* **86**, 1203–1214 (2015).
38. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
39. Harrison, P. J., Lyon, L., Sartorius, L. J., Burnet, P. W. & Lane, T. A. The group II metabotropic glutamate receptor 3 (mGluR3, mGlu3, GRM3): expression, function and involvement in schizophrenia. *J. Psychopharmacol.* **22**, 308–322 (2008).
40. Saini, S. M. et al. Meta-analysis supports GWAS-implicated link between GRM3 and schizophrenia risk. *Transl. Psychiatry* **7**, e1196 (2017).
41. Yang, X., Wang, G., Wang, Y. & Yue, X. Association of metabotropic glutamate receptor 3 gene polymorphisms with schizophrenia risk: evidence from a meta-analysis. *Neuropsychiatr. Dis. Treat.* **11**, 823–833 (2015).
42. Jia, W. et al. Metabotropic glutamate receptor 3 is associated with heroin dependence but not depression or schizophrenia in a Chinese population. *PLoS One* **9**, e87247 (2014).
43. Jablensky, A. et al. Polymorphisms associated with normal memory variation also affect memory impairment in schizophrenia. *Genes Brain Behav.* **10**, 410–417 (2011).
44. Baune, B. T. et al. Association between genetic variants of the metabotropic glutamate receptor 3 (GRM3) and cognitive set shifting in healthy individuals. *Genes Brain Behav.* **9**, 459–466 (2010).
45. Uchida, T. et al. A novel epidermal growth factor-like molecule containing two follistatin modules stimulates tyrosine phosphorylation of erbB-4 in MKN28 gastric cancer cells. *Biochem. Biophys. Res. Commun.* **266**, 593–602 (1999).
46. Kanemoto, N. et al. Expression of *TMEFF1* mRNA in the mouse central nervous system: precise examination and comparative studies of TMEFF1 and TMEFF2. *Brain Res. Mol. Brain Res.* **86**, 48–55 (2001).
47. Horie, M. et al. Identification and characterization of TMEFF2, a novel survival factor for hippocampal and mesencephalic neurons. *Genomics* **67**, 146–152 (2000).
48. Siegel, D. A., Davies, P., Dobrenis, K. & Huang, M. Tomoregulin-2 is found extensively in plaques in Alzheimer's disease brain. *J. Neurochem.* **98**, 34–44 (2006).
49. Lin, H. et al. Tomoregulin ectodomain shedding by proinflammatory cytokines. *Life Sci.* **73**, 1617–1627 (2003).
50. Psych, E. C. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).

Acknowledgements

The authors are grateful to C. Niswender, B. Stansley, and J. Conn at Vanderbilt University for critical input on portions of this manuscript. This study is supported by US NIH/NHGRI grants U01HG009086 (to Q. Wang, R.C., Q. Wei, Y.J., H.Y., X.Z., R.T., N.J.C., and B.L.), U24HG008956, and R01MH113362 (to J.S., N.J.C., and B.L.). The grant U01HG009086 supports the Vanderbilt Analysis Center for the Genome Sequencing Project (GSP) and U24 HG008956 supports the GSP Coordinating Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

B.L. conceived the overall design of the study, with Q. Wang and R.C. providing input. Q. Wang and R.C. implemented the algorithm and performed most of the analyses. F.C., Q. Wei, Y.J., H.Y., X.Z., and R.T. provided data integration and analyses. Z.W., J.S.S., C.L., E.H.C., and N.J.C. contributed to the interpretation of the results. Q. Wang, R.C., F.C., and B.L. wrote the manuscript, and all authors participated in the review and revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0382-7.

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

Correspondence and requests for materials should be addressed to B.L.

Journal peer review information: *Nature Neuroscience* thanks Vincent Mooser, Hon-Cheong So, and other anonymous reviewer(s) for their contribution to the peer review of this work.

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

Methods

Model description of iRIGS. We collected genes in the 2 Mb region centered at a GWAS index SNP as the candidates for that particular locus. L represents the number of GWAS loci, and we denoted a vector of genes with length L, each being from one of the L GWAS loci, as (X_1, …, X_L), and termed it a candidate risk gene set (CRGS). We denoted the corresponding genomic features for this CRGS as (D_1, …, D_L), in which D_l is a vector of genomic features collected for gene X_l, and let D be the genomics data for all candidates across all GWAS loci. We used N to denote the gene–gene network. Now the goal was to calculate $P(X_1, …, X_L|D,N)$ and to select a CRGS with a maximum PP, which was conditional on the collective genomic features on all genes in the L loci and on the network topology. Assuming that the genomics data of a gene only depend on the underlying gene, the following equation was applied:

$$P(X_1,\ldots,X_l|D,N)\propto P(D|X_1,\ldots,X_l)P(X_1,\ldots,X_L|N)=\prod\nolimits_{l=1}^{L}P(D_l|X_l)P(X_1,\ldots,X_L|N)$$

The first term represents the evidence embedded in genomics data, and the second term encodes the complex correlation of risk genes in a network. Since it is impossible to explicitly specify the correlation among the genes, we implicitly derived the complex correlation from the network, with the rationale that the functional convergence of risk genes, reflected in the perspective of networks, is that disease genes are more densely connected and are therefore more highly correlated. Specifically, we approximated it with one-dimensional conditional likelihoods; that is, $P(X_1,\ldots,X_L|N)\approx\prod\nolimits_{l=1}^{L}P(X_l|X_{-l},N)$, where X_{-l} is a vector of genes with the l-th gene removed. We can see that the joint PP can be approximated by one-dimensional pseudo-likelihoods as follows:

$$P(X_1,\ldots,X_l|D,N)\propto\prod\nolimits_{l=1}^{L}P(X_l|X_{-l},D,N)=\prod\nolimits_{l=1}^{L}P(D_l|X_l)P(X_l|X_{-l},N)$$

Next, the calculation was decomposed into a one-dimensional problem, evaluating one GWAS locus at a time. For each of the genes in locus l, the evidence came from the following two sources: support from the genomics data (that is, $P(D_l|X_l)$) and support from risk genes in other loci through networks (that is, $P(X_l|X_{-l},N)$). Suppose that all of the X_{-l} genes are risk genes, then, based on network topology, a gene at locus l that is closer to X_{-l} is more likely to be the risk gene compared with other candidates at the same locus. We do not know, however, which genes in other loci are risk genes; therefore we were not able to pre-specify risk genes X_{-l}. Conceptually, we employed a Gibbs sampling strategy to first sample a candidate risk gene from a given locus l based on the one-dimensional posterior, and then repeated the sampling across the remaining loci. We iterated the sampling process until the posterior distribution converged. Specifically, in each round of Gibbs sampling, we calculated the sampling frequency for each candidate gene. The frequency was compared with the last round, and if the sum of squares of frequency differences across all selected genes was smaller than a predefined threshold (1×10^{-4} was used in this study), the sampling procedure stopped. On the basis of this sampling, we were able to assess the confidence of candidates being risk genes. Theoretically, we cast iRIGS as a Bayesian model selection problem, with each candidate in a locus being a risk gene as a model. We also defined a null (background) model X_0 to represent that the candidate is a non-risk gene. The Bayesian model selection method calculates posterior odds of X_l over X_0; that is, $\frac{P(X_l\mid X_{-l},D,N)}{P(X_0\mid X_{-l},D,N)}=\frac{P(D_l\mid X_l)}{P(D_l\mid X_0)}\frac{P(X_l\mid X_{-l},N)}{P(X_0\mid X_{-l},N)}$, where $\frac{P(D_l\mid X_l)}{P(D_l\mid X_0)}$ is a Bayesian factor derived from multi-omics data and $\frac{P(X_l\mid X_{-l},N)}{P(X_0\mid X_{-l},N)}$ is a prior odds derived from the network. The prior odds reflect the network evidence supporting X_l, with the rationale that the prior odds are high when X_l is closer to X_{-l} in the network compared with X_0. The distance of X_l or X_0 to X_{-l} in the network was calculated using the random walk with restart algorithm (Supplementary Note). We collected seven genomic features to compute the Bayesian factor, including DNM, DE, DTS, and four sets of regulatory connections determined by DRE–promoter links from the Hi-C, capture Hi-C, and FANTOM5 data. We employed the Mahalanobis transformation[51] to decorrelate the integrated multidimensional data so that any supportive genomic features were properly incorporated (Supplementary Note). In implementation, we assumed that the PP of the null model $P(X_0|X_{-l},D,N)$ is invariant for all candidate genes and thus only calculated $P(X_l|X_{-l},D,N)$. The application of iRIGS to 108 SCZ loci with 7 different genomic features took ~2 h on an Intel Xeon E5 central processing unit with 2.40 GHz.

Gene set enrichment analysis. Gene sets from various sources that exhibited strong evidence of their involvement in SCZ were collected for gene set enrichment analysis (Table 1). These gene sets included the following: FMRP targets extracted from two previous studies[52,53]; PSD genes[17,54]; genes related to presynaptic proteins (PRP), PRAZ, and synaptic vesicles (SYV)[55]; the GABA$_A$ receptor complex[37]; CCS genes[56]; and targets of miR-137[16]. In addition to the primary SCZ functional categories, we collected a few autism spectrum disorder (ASD) gene sets for enrichment analysis owing to the pathophysiology shared between psychiatric disorders. These gene sets included the following: genes from the database AutDB[57]; evolutionarily constrained genes (ECG)[58]; essential genes[59]; genes from transmission and de novo association test (TADA)[60]; and targets of RBFOX1 (RNA binding protein, fox-1 homolog 1), a brain- and muscle-specific splicing factor[61].

We also compiled gene sets relevant to CNS phenotypes in mouse models[37]. We leveraged the phenotypic terms in MPO, a well-constructed vocabulary that unambiguously describes phenotypic observations[62], and gene–phenotype relationships in MGI[63] to extract CNS gene sets. First, we identified 2,066 descendant terms of the following two relevant terms of the highest level: nervous system phenotype, and behavior and neurological phenotype. Next, we downloaded all gene mutations of the mouse and their MPO annotations from MGI. Since the MPO was constructed in a hierarchical structure, we assigned genes annotated to a specific term to all its ancestry terms. We then mapped the mouse genes to human genes using Human and Mouse Homology Classes generated in MGI. We only kept the homology classes that contained unambiguously orthology relationships; that is, the classes consist of only a single mouse–human gene homolog pair. Finally, we obtained 278 terms that each contained at least 50 human genes.

DNM enrichment analysis. We collected the SCZ DNM data from multiple previous studies[11,36,64,65], in which exome sequencing was performed on parent–proband trios, and, in some cases, with an unaffected sibling. In total, the sequenced cohort consisted of 973 trios and 84 unaffected siblings. We annotated the DNMs by ANNOVAR[12] and extracted the following two classes of DNMs: (1) loss of function (LoF) mutations, including nonsense, splicing, and frame shift, and (2) missense mutations. Twelve bioinformatics tools were utilized to determine the deleteriousness of LoF and missense DNMs in ANNOVAR, and we assigned a DScore for each missense mutation, defined as the number of deleteriousness predictions out of 12 prediction algorithms from ANNOVAR. We only focused on the pdDNMs, defined as LoF and missense DNMs with a DScore of >3. Accordingly, the control set we used included pdDNMs identified in all the control samples collected by denovo-db.

Gene expression analysis. For the DE analysis, we downloaded data from the CommondMind Consortium (https://www.synapse.org/#!Synapse:syn5609493)[13], in which RNA sequencing was performed on post-mortem dorsolateral prefrontal cortex region samples from 258 subjects with SCZ and from 279 controls. We then employed the Wilcoxon rank-sum test to see whether the iRIGS-predicted risk genes carry lower P values compared to the background.

For the tissue-specificity investigation, we used gene expression data from the GTEx release V6[66]. We downloaded gene RPKM (reads per kilobase of transcript per million mapped reads) dataset from the GTEx portal (https://www.gtexportal.org/home/datasets), covering ~50 tissues. We adopted the Jensen–Shannon divergence[67] to measure the tissue specificity of each gene in each tissue (Supplementary Note).

For the brain developmental stage-specificity investigation, we downloaded the RNA sequencing data of developing human brains from BrainSpan[15], and calculated the average expression of HRGs in all brain regions at each of the developmental stages (http://help.brain-map.org/display/devhumanbrain/Documentation). We used the log$_2$(RPKM) as the expression level of genes.

DRE–promoter link collection. We collected DRE–promoter links from multiple sources. One recent study[9] inferred chromosome contact by constructing Hi-C libraries for two major regions, the cortical and subcortical plate and the germinal zone, of the human cerebral cortex. The predicted DRE–promoter links listed in supplementary tables 22 and 23 of that study were downloaded, and in total we obtained 221,069 and 228,323 links for the cortical and subcortical plate and the germinal zone, respectively. We also collected DRE–promoter links inferred from two other studies[8,10]. One is a capture Hi-C study of the cell line GM12878[10]. We obtained 1,618,000 DRE–promoter links predicted for GM12878 from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2323/. The other dataset we used is from the FANTOM5 project[8], in which the cap analysis of gene expression technology was employed to infer the enhancer–promoter links across multiple human tissues. We downloaded the FANTOM5 data from http://enhancer.binf.ku.dk/presets/ and obtained 66,899 enhancer–promoter links.

Construction of drug–target network. We collected drug information from the DrugBank database (v.4.3)[68] and the Therapeutic Target Database (accessed on December 2016)[69]. All chemical structures from these databases were prepared using the Open Babel toolkit (v.2.3.2)[70]. We assembled bioactivity data for drug–protein interactions collected from the following three publicly available databases: ChEMBL (v.21)[71]; BindingDB (data accessed on December 2016)[72]; and the IUPHAR/BPS Guide to Pharmacology (data accessed on December 2016)[73]. To improve data quality, we only pooled the biophysical drug–protein interactions with the numeric bioactivity value using the following four criteria: (1) K_i (inhibition constant), K_d (dissociation constant), IC$_{50}$ (half-maximum inhibitory concentration) or EC$_{50}$ (half-maximum effective concentration) values of ≤10 μM; (2) the target protein can be represented by a unique UniProt accession number; (3) the target protein was marked as 'reviewed' in the UniProt database[74]; and (4) the target protein is from *Homo sapiens*. A fixed length (25 hash characters) generated from chemical SMILES by OpenBabel[70] was used to encode each drug. All duplicated drugs were removed according to their 25 hash characters. Drugs were grouped using anatomical therapeutic chemical (ATC) classification system codes collected from DrugBank[68]. We defined antineoplastic drugs based on the first-class of ATC code, such as [N] for 'nervous system' drugs.

Statistical analyses. For the gene set enrichment, DNM enrichment, and drug–target enrichment analyses, we adopted the one-sided Fisher's exact test. For PP comparisons, spatiotemporal expression analyses, and DRE–promoter link comparisons, we adopted the one-sided Wilcoxon rank sum test.

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

Data availability

All the data used in this study are from public resources that are specified in the Methods and the Supplementary Note.

Code availability

The source code and the companying genomics datasets used in this study are available at https://www.vumc.org/cgg.

References

51. Härdle, W. & Simar, L. *Applied Multivariate Statistical Analysis* (Springer, 2007).
52. Ascano, M. Jr et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**, 382–386 (2012).
53. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
54. Bayes, A. et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* **14**, 19–21 (2011).
55. Pirooznia, M. et al. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28**, 897–899 (2012).
56. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
57. Basu, S. N., Kollu, R. & Banerjee-Basu, S. AutDB: a gene reference resource for autism research. *Nucleic Acids Res.* **37**, D832–D836 (2009).
58. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
59. Ji, X., Kember, R. L., Brown, C. D. & Bucan, M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proc. Natl Acad. Sci. USA* **113**, 15054–15059 (2016).
60. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
61. Weyn-Vanhentenryck, S. M. et al. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).
62. Smith, C. L., Goldsmith, C.-A. W. & Eppig, J. T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6**, R7 (2005).
63. Blake, J. A. et al. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45**, D723–D729 (2017).
64. Girard, S. L. et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
65. Xu, B. et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
66. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
67. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
68. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
69. Yang, H. et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* **44**, D1069–D1074 (2016).
70. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
71. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
72. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).
73. Pawson, A. J. et al. The IUPHAR/BPS guide to pharmacology: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.* **42**, D1098–D1106 (2014).
74. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

nature research

Corresponding author(s): Bingshan Li

Last updated by author(s): Mar 8, 2019

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: All chemical structures from DrugBank database (version 4.3) and Therapeutic Target Database (accessed on December, 2016) databases were prepared by the Open Babel toolkit (version 2.3.2).

Data analysis: R version 3.2.2 was used for statistical analysis. We also developed new codes in R this study and the source code availability is in the main manuscript under "Code availability". We used ANNOVAR (version 2016Feb01) for annotation.

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

The source code and the companying genomics datasets used in this study are available at https://www.vumc.org/cgg.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used public data, and the sample sizes were the same as the original data and vary from dataset to dataset.
Data exclusions	No data were excluded.
Replication	We used different datasets from public data to test the same hypotheses and the results were replicated in this way.
Randomization	This study is the development of a new statistical and computational method and randomization does not apply here
Blinding	This study is the development of a new statistical and computational method and randomization does not apply here

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging